Evaluation of functioning in workers with whiplash-associated disorders and back pain

Trippolini, Maurizio

Printing of this thesis was generously supported by donations received from:

Rehaklinik Bellikon, Switzerland
Verein IG Ergonomie, Swiss Association of Rehabilitation
Centrum voor Revalidatie, UMCG
Ontwikkelcentrum Pijnrevalidatie – Centrum voor Revalidatie, UMCG
Stichting Beatrixoord Noord-Nederland
WorkWell Inc., Duluth, USA

Cover picture Margriet Barends, Ulrum. www.margrietbarends.com
Layout Renate Siebes, Proefschrift.nu
Printed by Ridderprint, Ridderkerk

ISBN 978-90-367-7162-7

rijksuniversiteit
groningen

Evaluation of functioning

in workers with whiplash-associated disorders and back pain

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. E. Sterken
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

maandag 8 september 2014 om 16.15 uur

door

Maurizio Alen Trippolini

geboren op 26 augustus 1975
te Samedan, Zwitserland

Promotores	Prof. dr. M.F. Reneman
	Prof. dr. P.U. Dijkstra
	Prof. dr. J.H.B. Geertzen
Beoordelingscommissie	Prof. dr. U. Bultmann
	Prof. dr. B.W. Koes
	Prof. dr. R.J.E.M. Smeets

Paranimfen Beatrice Jansen
Haitze de Vries

Für Corinne
Emma, Sofia & Teo

Contents

Chapter 1

General introduction

DEFINITION AND ETIOLOGY

The Québec Task Force (QTF) has defined whiplash as "an acceleration-deceleration mechanism of energy transferred to the neck that results in soft tissue injury that may lead to a variety of clinical manifestations including neck pain and its associated symptoms" [1]. These symptoms may include upper limb pain, jaw pain, paraesthesia in the upper limbs, dizziness, headache, fatigue, nausea, concentration deficits, psychological distress, anxiety and other complaints [1]. The QTF coined the term whiplash-associated disorders (WAD) to describe the injury and its related symptoms. Pain in the anatomical region of the neck is the main symptom of WAD (Figure 1.1).

In this thesis, the term WAD[1] will be used. WAD may occur following after rear end or side-impact motor vehicle collisions, but can also be the result of work, sports or other mishaps which produce indirect cervical trauma. WAD is classified based on clinical observations into 1 of 5 categories (grade 0-IV) (Table 1.1).

Although several adaptations of the original QTF classification have been proposed [2-4], the QTF 1995 classification is still the most frequently used in research and clinical practice with WAD. Grade I and II of the QTF classification refer to neck complaints and musculoskeletal signs such as reduced range of motion and point tenderness and represent more than 90% of patients with WAD [5-7]. The etiology of WAD likely combines physical and psycho-social factors, nevertheless the pathophysiology is not well understood [8]. Hence, WAD has been described as a "systemic illness" or "functional somatic syndrome" with no single etiological factor [9-11]. As such, WAD can be understood within the biopsychosocial model [12] which has been re-conceptualized into a model for neck disorders by the Task Force on Neck Pain and its Associated Disorders [4]. The biopsychosocial model highlights health, illness and disability as the product of a combination of factors, including the biology of the individual (e.g., genetic predispositions, chemical imbalances), behavioral factors (e.g., lifestyle, stress-reaction, health beliefs) and social conditions (e.g., cultural influences, family relationships, social support) [13].

Epidemiology, prognosis and burden to society

Despite the ongoing debate about whether or not WAD can be viewed as a valid disease definition [14] the annual incidence of whiplash injuries per year is substantial in many Western countries: Sweden and Germany 1-3.2/1000, Canada and United States 3.5-7/1000, United Kingdom 4.2/1000, The Netherlands 1.9-3.3/1000 and 2.8/1000 in Switzerland [2,15,16]. Prevalence is difficult to estimate and estimates are usually the 0.4-2% range for

1 Other terms used in the scientific literature to describe whiplash are: whiplash trauma, whiplash injury, neck sprain, neck injury, whiplash disorder, and chronic or late whiplash syndrome.

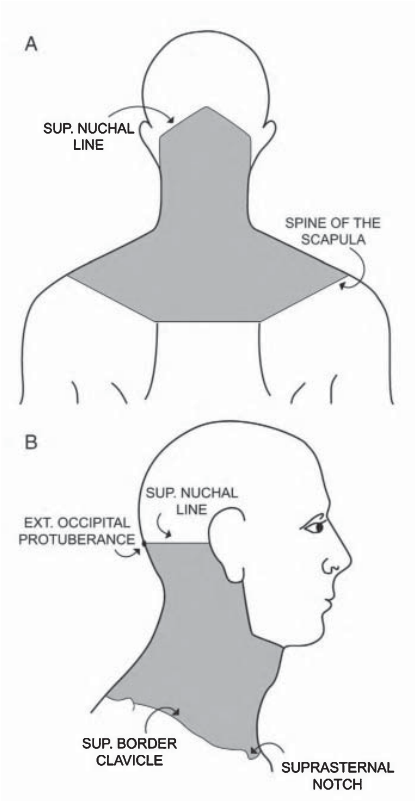


Figure 1.1 The anatomical region of the neck from the back (A) and the side (B) as defined by The Neck Pain Task Force and its Associated Disorders. Reprinted from Guzman J et al. A new conceptual model for neck pain. Journal of Manipulative Physiological Therapeutics. (2008) 32; 2S. © Copyright: Lippincott Williams & Wilkins.

the general population [2,17]. Although the prognosis of WAD is generally favorable, with a self-reported recovery rate of 40-60% within the first 12 months, a considerable number of individuals with WAD still report symptoms and disability 12 months after the injury [18,19]. Established prognostic factors include pain intensity post-injury and self-reported disability after the injury [7]. Psychosocial factors such as high fear of movement, low self-efficacy beliefs, low expectation of recovery, high levels of pain catastrophizing, poor coping strategies and depression predict poor recovery [18,20-22]. Studying the prognosis of whiplash is complicated and the validity of previous studies has been limited by small sample size, inclusion of patients more than 6 months after injury onset, short follow-up periods (less than 6 months), loss to follow-up, unblinded outcome assessors and lack of statistical adjustment for important covariates [23].

Table 1.1 The Québec Task Force (QTF) Classification of Whiplash-Associated Disorders

Grade	Clinical presentation
0	No complain about the neck No physical signs
I	Neck complaint of pain, stiffness or tenderness only No physical sign(s)
II	Neck complaint AND musculoskeletal sign(s) [a]
IV	Neck complaint AND neurological sign(s) [b]
V	Neck complaint AND fracture or dislocation

[a] Musculoskeletal signs include decreased range of motion and point tenderness.
[b] Neurological signs include decreased or absent deep tendon reflexes, muscle weakness, sensory deficits (dizziness, tinnitus, headache, memory loss, dysphagia and temporomandibular joint pain).
QTF classification of whiplash injuries, adapted from the Swedish Task Force. Reprinted from Jansen GB et al. Whiplash injuries: diagnosis and early management. The Swedish Society of Medicine and the Whiplash Commission Medical Task Force. Eur Spine J. (2008); 17 Suppl 3:S355-417. © Copyright: Springer.

Over the last decade the number of patients in Switzerland who claimed a whiplash injury varied, with approximately 21,000-27,000 new cases per year [24]. Of these patients, 5-10% have had a delayed recovery and developed a chronic course of WAD [25]. Delayed recovery of WAD places a substantial burden on the individual and society due to long-term sickness absence and work disability [26]. The costs of WAD have fluctuated greatly over recent decades. By the late 1990s, the costs for work loss and health care due associated to WAD in Switzerland have increased five-fold compared to the 1980s [25]. Whiplash injuries were responsible for 50% of the costs paid by third party for injuries due to car accidents [25]. In 2002, Switzerland had the highest expenditure per claim in Europe, an average of 35,000 euros compared 9,000 euros in the rest of Europe [26]. Nevertheless, data from Swiss Central Office for Statistics in Accident Insurance shows that since 2003 the total costs of health care and workers compensation caused by whiplash injuries has reduced from 497 million Swiss francs (408 million euros) to 152 million Swiss francs (125 million euros) in 2012 [24]. Whether these reductions in costs are due to the effect of a more intensive support of patients (e.g. by case management), early screening including interdisciplinary WAD assessments of patients at risk for chronic course, better adherence to treatment guidelines by health care providers, changes in legislation or a combination of these factors is unknown.

Role of the context in developing WAD

Until the 1990s, many studies searching for a discriminating biomedical factor of WAD were published; in the last two decades, the role of contextual factors, such as the social security system, and cultural aspects on the incidence and course of WAD has increasingly been acknowledged [27,28]. Although comparisons *between* countries showed large differences in the incidence of WAD, the cultural factors such as mother language, being a native-speaker, which may influence the course of WAD *within* the same country, have rarely been reported. Whether or not compensation of the consequences of the whiplash injury plays an important role in the chronic course of WAD is still under debate [29-31]. Perceived injustice by patients with WAD related to the instigator of the accident or the insurer may play a relevant role in the chronic course of WAD [32]. Although the vast majority of individuals with WAD reported no symptoms prior to the whiplash injury [9], the results of systematic reviews and large population based studies indicate that aspects of pre-injury health status such as neck pain, anxiety and depression influence the incidence of WAD [19,33,34].

Strategies to manage the burden

Despite the large body of research and three large WAD Task Forces, very few interventions have proved effective in the treatment of acute or chronic WAD [1,35]. Interventions can be classified in two groups: early management of WAD (4-12 weeks after injury) and management of patients with delayed recovery after whiplash injury (>12 weeks). In the early management of WAD, information about the usually favorable course and advice towards rapid return to normal activity appears effective [2,36]. The use of a cervical collar is not recommended. Moreover, any pharmacological or other treatment should be regularly followed to avoid adverse effects such as addiction [35]. For patients with persistent symptoms of WAD (>12 weeks after injury) evidence for effectiveness of multidisciplinary interventions is lacking. While for patients with chronic WAD some studies have shown promising results [37,38], other population based studies have not [39]. The growing body of evidence suggests that type, intensity, and timing of health care delivery are strongly and independently associated with time to recovery [40]. Finally, it is apparent that similar to other musculoskeletal disorders, one size does not fit all when targeting effective treatments in chronic WAD [41]. Without a doubt, the most important achievement would be to prevent chronic WAD. Innovative trials, which would use a more social-policy or public health approach, e.g., media campaigns similar to the ones performed in back pain, are warranted [42,43].

Measurement of functioning, work capacity and disability determination

This thesis is embedded in the conceptual framework of the International Classification of Functioning, Disability and Health (ICF) [44]. Within the ICF framework, human functioning is classified into several factors based on the biopsychosocial model. Functioning is constituted by body structure and functions as well as activities and participation (Figure 1.2). Additionally, functioning is influenced by environmental and personal factors. In this thesis, functioning was measured at the activity level with functional capacity tests and self-reported measures. Functioning is a perquisite for work capacity.

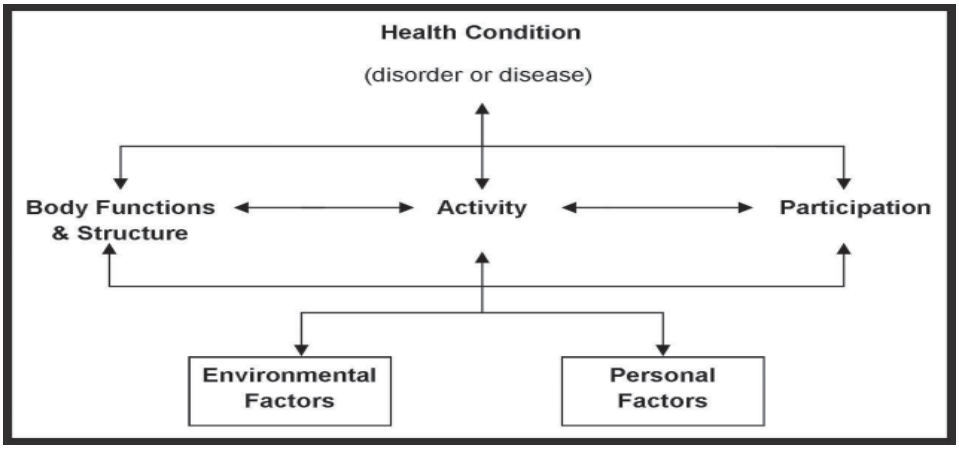


Figure 1.2 International Classification of Functioning, Disability and Health [53].

The traditional disease model describes a causal pathway from diagnosis to disability [45]. The basis of the biopsychosocial model in chronic disorders is that the relationship between diagnosis and disability are less evident. In the 1960s, Dave Mechanic stated that in chronic disorders "disease and disability may vary independently" [46], i.e., in many chronic disorders the medical diagnosis and the objectively measured functional limitation are weakly correlated. Surprisingly, although there is agreement in medicine on the biopsychosocial model and Mechanic's statements, work capacity certificates are still based on medical doctors' biomedical findings, very few of which are associated with return to work [47,48]. It has been increasingly acknowledged that patients with nonspecific illnesses, such as WAD, require a functional evaluation, as medical diagnosis and structural findings alone are inappropriate to determine work capacity [49,50]. Several countries such as the UK, the Netherlands, Denmark and Norway have changed their policies of disability determination from a diagnosis based approach to an evaluation of the remaining functional capacity of

a person matched to the job requirements [47,50,51]. Similarly, Swiss legislation requires the physician to judge the "inability to work" primarily by the extent of the functional loss concerning the job demands of the previous work and not by diagnosis [49]. According to the guidelines of the International Labor Organization sick or disabled persons should be assessed comprehensively to avoid over- or underestimation of work (dis)ability [52].

Self-reported functioning

The patient's perceived functional self efficacy (SE) level was proposed by Bandura in the 1970s as a relevant psychosocial factor that may influence the performance of an individual [54]. Perceived SE refers to the individual's beliefs about their own competence or ability [54]. SE beliefs may influence patient behavior, e.g., the ability to overcome negative experiences. It has been suggested that SE is more closely related to work disability than actual physical abilities [55]. Assessment of SE, usually measured by questionnaires, plays an important role in predicting physical performance and health outcomes [56-58]. Patients with low SE may have a low perceived functional ability, which may affect their work functioning. However, the utility of questionnaires is often limited by literacy level and linguistic abilities [59], forming a barrier for targeting rehabilitative interventions in patients with low literacy levels [60]. One approach to improve the comprehension of the questionnaire is to inform the patient through depicted activities in combination with each item [61]. The Spinal Function Sort (SFS), published in English in 1989 [62], is a picture-based generic tool consisting of 50 depicted tasks (example: Figure 1.3). These tasks reflect a wide range of daily living or vocational activities that involve the spine. The depicted activities are graded from light to heavy material handling, so that scores can be compared to the level of physical workload according to the Dictionary of Occupational Titles [63]. The SFS claims to measure perceived functional ability based on the SE theory. The SFS has proven high practicality in different rehabilitation settings [64-68] and is used in addition to functional capacity evaluations [69]. The reliability and validity of the SFS have been reported [70-72] but, to the best of our knowledge, no German or French versions have been properly cross-culturally adapted and translated. In addition, the SFS measurement properties have not been studied in patients with WAD.

Self-reported versus performance-based measures of functioning within the context of a social-security system

Relying on patient self-reports as a means to assess functional capacity and workloads may have some limitations when compared with measurement of functional capacity and workloads because it seems to result in an underestimation of functional capacity and an overestimation of workload [73-76]. When asking about physical complaints within a

Figure 1.3 Item 14 of the Spinal Function Sort (SFS) questionnaire: Lift a 10 kg milk crate from the floor to eye-level. Instructions: *This is a test of your current ability to perform work tasks. Look at each drawing and read the description. On the separate answer sheet, indicate your current level of ability to perform the task* ("able", "restricted" or "unable"). © Copyright: PACT 1989. All rights reserved.

compensation environment there is room for misattribution, and over reporting of accident-caused complaints [77]. Most (prognostic) studies on WAD used self-reported measures or patient-record data [19]. Self-reported measures, while meant to measure perceptions, are also susceptible to information bias due to compensation environments, social desirability, dissimulation or response style [78,79]. Additionally, there appears to be a weak to moderate association between self-reported and objectively measured function in patients with chronic pain [55,80,81]. Performance and behavioral tests may overcome these shortcomings and provide a more comprehensive picture of the current (work-related) functioning of patients within the biospsychosocial model of chronic pain.

Functional Capacity Evaluation

One method to determine function is the Functional Capacity Evaluation (FCE). FCE consists of a battery of standardized tests to evaluate a person's functional capacity and safe ability to work [82] (two examples of FCE tests are displayed in Figure 1.4). FCE systems were developed in the mid 1970s in the USA [83]. The workers compensation law required physicians to assess patients' ability to work beyond providing medical information and venture into the realm

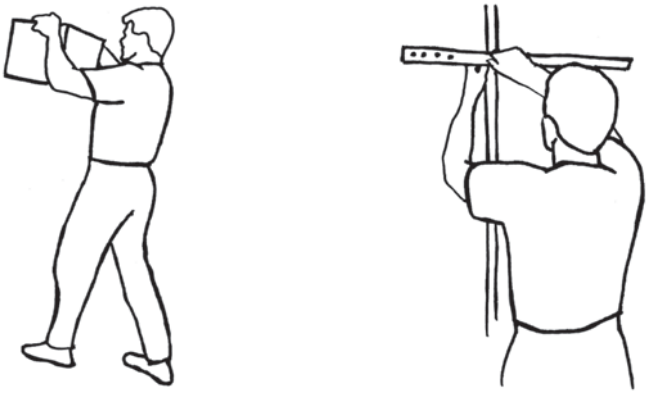


Figure 1.4 Two FCE tests involving the cervical region and the upper extremities: lifting waist to overhead (left image), and overhead working (right image).

of return-to-work specifics [83]. Physicians turned to their functional counterparts, physical and occupational therapists, to provide information about work function.

In Switzerland, the standard two-half day FCE as described by the Isernhagen Work System (now, WorkWell FCE) [82] is administered by a physician and a physical or occupational therapist specializing in work rehabilitation [69]. The FCE includes a summary of medical records, a physical exam, a comprehensive assessment of psychosocial risk factors measured by a interview or questionnaire, an assessment of perceived work ability with the SFS, a measurement of physical capacity with FCE tests and additional job specific activities, observation-based ratings of pain behavior and physical effort, and determination of consistency during testing (Table 1.2. FCE format).

Table 1.2 FCE format

- History and summary of medical records
- Physical exam
- Self-reported measures (e.g. pain, psychosocial risk factors, perceived work ability)
- FCE tests and job specific activities
  - Pain behavior
  - Level of effort
  - Consistency of test results
- Matching physical demand level and functional capacity
- Written report with recommendations

Adapted from Isernhagen SJ. Introduction to Functional Capacity Evaluation. In Elizabeth Genovese, Jill S. Galper. Ed. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations. Chicago IL: American Medical Association, 2009:1-18.

Additionally, physical capacity as determined by FCE tests is compared with the required physical job demands of the patient's occupation [82]. Critical job demands are assessed by a job analysis, which is based on an interview with the patient, the employer, direct observation on site, or existing job descriptions. This information is merged in a FCE report and recommendations for participation in work are made, including work-adaptations, and fitness for work determination for the current job or for other jobs (unemployment). When the results of an FCE indicate that a worker's functional capacity does not match the physical demands of the job, a rehabilitation program can be proposed to enhance the worker's ability to return to work [66,84]. Data from FCE improve the quality of the medical work capacity determination process and therefore facilitate the return-to-work process or prelude case closure [85,86].

FCE tests applied in the studies for this thesis

The FCE tests applied in this thesis are: handgrip strength (left and right), lifting floor to waist, lifting waist to overhead, short two-handed carry, long right- and left-handed carry, overhead work, repetitive reaching (left to right and right to left [87], 50 meter walking test [88] and a 3 minute step test [89]. Patients were briefly instructed on how to perform each test; the rater first gave a short demonstration of each test. Patients were then asked to perform the tests to their maximum ability. Weights lifted were gradually increased according to a participant's performance, using weights of 2.5 and 5 kilograms. The duration of the FCE tests used in the WAD Assessment is approximately 60 min.

The context of the FCE tests within this thesis

The FCE tests studied in this thesis are embedded in an interdisciplinary assessment for patients with WAD, which was developed in 2005 at the Rehaklinik Bellikon. A rehabilitation physician conducts a review of the medical history and a physical examination; a physiotherapist then administers FCE tests. After determination of eligibility, patients complete questionnaires and carry out FCE tests. The FCE tests are followed by a brief intervention aimed to enhance recovery. The intervention usually includes a therapy trial with a combination of manual therapy, exercises, ergonomic advice on manual material handling, education on how to stay active, deal with symptoms and flair-ups, and advice for home exercises. The interdisciplinary rehabilitation assessment ends with a face-to-face discussion with the patient about medication, treatment options and possible strategies to facilitate recovery and return to work. Fitness-for-work certificates or work capacity settlements are explicitly not part of this interdisciplinary assessment. A written report about the content and the results of the assessment is sent within two weeks to the referring insurance physician, general practitioner

and/or case manager. The studies regarding WAD this thesis used are embedded within the WAD Assessment of the rehabilitation clinic in Bellikon.

Patients with WAD participating in this thesis were workers who were insured by the Swiss Accident Insurance Fund (SUVA). SUVA is the largest state owned accident insurance fund in Switzerland and covers occupational and non-occupational injuries for employed individuals, mainly in labor industries, and unemployed job-seeking persons [90]. Injured persons receive compensation up to a maximum of 80% of their previous salary, and medical and vocational assistance. If health status is stabilized but disabilities remain, long-term invalidity pensions are refunded by SUVA and the invalidity insurance.

Physical effort determination during FCE tests

In FCE tests maximal effort of the patient required to reach valid results [82]. If FCE test results are hampered by "submaximal effort" they may lead to ineffective treatment, wrong disability classifications and inappropriate compensation [91]. Despite several measures that are used to measure physical effort, such as force ratios, electromyography and motion ratios, very few are thoroughly validated in patients with chronic pain [91-93]. The FCE tests performed are based on the kinesiophysical approach proposed by Isernhagen [82]. Within this approach, the level of physical demand is determined by the rater using observational criteria indicative for physical effort (Table 1.3) [83,94]. Based on a scale with observational criteria the physical demands for material handling tasks are classified into three categories: "light-to moderate," "heavy," or "maximal." Observational criteria for postural tolerance tests and ambulation tests are rated on a scale from "no or slight functional problem/limitation," "some functional problem/limitation" to "substantial functional problem/limitation." FCE testing is terminated for two reasons: a) the participant stopped at the "submaximal" level because, for example, pain, fear or other patient-reported reasons; or b) the rater stops when based on observational criteria maximal safe function is observed; heart rate exceeded 85% of the age-related maximum (220 minus age of participant); or a predefined time limit of the FCE test was reached. "Submaximal" effort is assumed if the patient stops the FCE test before the criteria indicative of maximum capacity are not observed. Although some narrative reviews have been performed in the field of physical effort determination [91-93], to date there is no empirical evidence on performance tests, which claim to measure "submaximal" performance.

Measurement properties of FCE tests

Many of the available FCE systems have been criticized in the past for not being rigorously analyzed according to required measurement properties such as reliability, validity and

Table 1.3 Observational criteria for determination of physical effort during material handling tests

Criteria	Light to moderate	Heavy	Maximum
Muscle recruitment			
Prime movers	Normal recruitment	Bulging	Bulging
Accessory muscles	No or only slight muscle recruitment	Distinct recruitment	Bulging
Base of support	Natural stance	Distinctly increased	Very wide base
Posture	No or only slight counterbalance in extension	Distinctly increased counterbalance	Substantial counterbalance
Heart rate and respiration	No or minimal increases in heart rate and respiration	Distinct increases in heart rate and respiration	Substantial increases in heart rate and respiration
Control and safety	Smooth movements	Increasingly controlled movement; might begin to use momentum; execution with difficulty but not yet at the limit	Still safe but unable to maintain control with the addition of any more weight
Pace	Moderate/comfortable pace	Distinctly slower; very deliberate movements	Very slow (an increased pace would affect stability and control)

Table 1.3 is used with permission from Verein IG Ergonomie, Swiss Association of Rehabilitation [94].

responsiveness [95-97]. Nevertheless, in the last decade several doctoral theses on the measurement properties of FCE tests have been written [98-107]. Among the available FCE systems the Isernhagen FCE (now WorkWell) was found to scientifically be the most thoroughly evaluated [108,109]. FCE tests were validated mainly in patients with low back pain [110-113], in healthy populations [114,115], and patients with osteoarthritis [116,117]. However, FCE was not tested in patients with WAD.

Gaps in FCE research

Despite the increased number of publications on FCE tests, it has been highlighted that substantial gaps in knowledge remain [118]. The methodological quality of the studies still shows large room for improvement. For example, it is unknown which tests are able to reliably detect "submaximal" effort when "maximal effort" is required during physical performance testing. Also, the reliability of the criteria for physical effort determination, which should help to interpret FCE tests, is hampered. Criteria for physical effort determination have not been validated for postural tolerance and ambulation tests. Also, criteria for physical effort

determination were not validated in representative sample of raters in a health care setting. In the majority of studies, FCE tests results were not accompanied by information regarding whether or not these tests results were performed with "submaximal" effort, limiting the interpretation of test results. For example, test-retest reliability studies were often performed on healthy workers [87] or patients with low back pain [110,119,120]. Blinding procedures for patients and raters are not reported in reliability studies [97,108,109]. Moreover, measures of agreement for clinical interpretation of FCE test results are seldom reported [97,108,109]. The few hypotheses driven construct validation studies on FCE tests performed in clinical populations had small sample sizes [118]. Differences in outcomes of FCE tests for different population groups have not been investigated, although there is growing cultural diversity within most European countries. Only one study compared FCE results of populations from three countries, and reported large unexplained differences between clinical populations [109]. Studies on the prognostic capacity of FCE tests did not include multiple time point measurements and did not use continuous outcome measures with established confounders. Lastly, the vast majority of studies on FCE were performed by the same Dutch and Canadian research groups and have yet to be replicated elsewhere [118].

MAIN RESEARCH QUESTIONS

Based on the lacking knowledge and insights described above, the main research questions of this thesis are:

Determination of physical effort

- Which instruments can detect "submaximal" physical and functional capacity when "maximal capacity" is requested?
- What is the inter- and intra-rater reliability of two observational scales, which are used to determine level of physical effort during FCE tests?

Functional capacity tests in patients with WAD

- What is the test-retest reliability and safety of FCE tests in patients with WAD?
- What is the construct validity of FCE tests in patients with WAD?
- What are the differences in FCE tests in patients with WAD with different cultural backgrounds?
- What is the predictive validity of FCE tests for future work capacity in patients with WAD?

Perceived functional ability

- What are the measurement properties of a picture-based questionnaire, which claims to measure perceived functional ability?

OUTLINE OF THE THESIS

Determination of physical effort

In *Chapter 2*, a systematic review is described, investigating which instruments detect submaximal physical and functional capacity when maximal capacity is requested. Knowledge about whether physical capacity is performed maximally or submaximally (i.e., not with full effort) may be essential for the appropriate classification of the allocation of rehabilitation interventions or disability settlement. The purpose of this review is to identify validated instruments that are sensitive to levels of performance and practical for use in clinical settings.

In *Chapter 3*, a reliability study is described, which measures the inter- and intra tester reliability of scales using observational criteria that claim to determine physical effort of FCE tests. The results will give a broader understanding of the accuracy of two different scales which are used by a representative sample of clinicians on a wide array of FCE tests such as material handling, postural tolerance, or ambulation tasks. Moreover, whether reliability of observer changes when measured twice within 10 months is reported.

Functional capacity tests in patients with WAD

The study in *Chapter 4*, addresses the test-retest reliability and safety of FCE tests in patients with WAD. This study will provide insight into the accuracy of the tests when applied by clinicians to patients with WAD. Moreover, knowledge about adverse effects of FCE testing of patients with WAD will become available.

In *Chapter 5*, a cross-sectional study describes the construct validity of FCE tests in patients with WAD. The aim of this study is to compare FCE tests with reference measures based on predefined hypotheses. The results will allow a better understanding of what FCE tests claim to measure in patients with WAD within different language groups (i.e., cultural backgrounds).

In *Chapter 6*, a longitudinal study is described in which the ability of FCE tests to predict future work capacity in patients with WAD is tested. A predictive model is developed on which future work status can be predicted.

Perceived functional ability

In *Chapter 7* and *8*, two studies are described evaluating the measurement properties of a picture-based questionnaire of perceived functional ability, the Spinal Function Sort (SFS). In *Chapter 7*, test-retest reliability, structural, construct and prognostic validity in patients with WAD are tested. In *Chapter 8*, the cross-cultural translation process, the reliability, the

agreement, and validity of the German and French version of the SFS in patients with back pain are described and tested.

General discussion

In *Chapter 9*, the findings of this thesis are integrated and reflected upon. Methodological considerations and implications of the findings for clinical practice are discussed. Recommendations for future research and final conclusions are made.

Appendix 1.1 Description of the main terms

Term	Description
Capacity	The highest probable level of functioning that a person can reach in a domain at a given moment in a standardized environment [121].
Criteria for physical effort determination	Signs that indicate the level of capacity of a person during activities by using observational criteria (e.g., posture, muscle recruitment, heart rate increase, control and safety).
Evaluation	A systematic approach including observation, reasoning, and conclusion. Going beyond monitoring and recording, the evaluation process implies an outcome statement that is explanatory, as well as an objective measurement. Evaluation is often used interchangeably with the term assessment.
Functional Capacity Evaluation (FCE)	A FCE is an evaluation of capacity of activities and is used to make recommendations for participation in work while considering the person's body functions and structures, environmental factors, personal factors and health status [121].
FCE rater	A FCE rater is the person which acts as evaluator during the FCE tests. The rater has had a formal training in FCE tests. He or she instructs the patient how to perform the tests. The rater observes the patient, measures parameters of the test (e.g., weight lifted, distance, etc.) and determines the level of physical effort on the basis of observational criteria. The term rater may be used interchangeably with the terms evaluator, assessor, observer and tester.
FCE test	A FCE test is a standardized, work-related item within a FCE. It is a clearly described measurement procedure quantified by counting weight, distance, time, repetitions etc. Physical demands are classified by means of criteria for physical effort determination.
Injury	Damage or harm done to or suffered by a person or a thing.
Maximal effort	The highest safe ability of a person during a FCE test [82]. Maximal effort was assumed when a FCE rater observed sufficient physical effort determination criteria indicative of safe maximal effort.
Measurement properties	The term describes the quality of an instrument to measure a construct. An example for a measurement property is reliability. Many synonyms and definitions are used for the same measurement properties. Measurement properties are evaluated by using methods of psychometrics or/and clinimetrics.

Appendix 1.1 continues on next page

Appendix 1.1 *Continued*

Term	Description
Patient	The term *patient* refers to the individual who is performing the FCE tests. The term *patient* can be used interchangeably with *evaluee, client* or *claimant*. The term *patient* was used in the studies of this thesis.
Performance	Describes "what a person does in the current environment" [121].
Perceived functional ability	The belief of individuals about their capabilities to perform work tasks.
Safety	Safety is a situation in which, given the known characteristics of the person, the procedure should not be expected to lead to injury.
Submaximal effort	Is less than a maximal level of functioning on the physical or activity level that a person may reach in a domain at a given moment in a standardized environment. Translated to a FCE test procedure, submaximal effort means that the patient stops a FCE test before the criteria indicative of a maximum capacity are observed.
Test	A standardized measurement procedure.
Whiplash associated disorders (WAD)	A variety of clinical manifestations, which are associated with a whiplash injury. WAD is classified according the Québec Task Force (QTF) classification (Table 1.1) [1].
Whiplash injury	Injury or functional impairment that occurs in association with a whiplash trauma [2].
Whiplash trauma	An acceleration-deceleration mechanism by which energy is transferred to the head and the neck without the existence of direct trauma to the head and the neck [2].
Work capacity (WC)	From the insurance perspective, WC is the proportion workability of the actual pre-injury work usually determined by the general practitioner, insurance or occupational physician. WC is expressed in a percentage (0-100%) of pre-injury work. WC can be translated into days or hours of modified work.

REFERENCES

1. Spitzer WO, Skovron ML, Salmi LR, Cassidy JD, Duranceau J, Suissa S, et al. Scientific monograph of the Quebec Task Force on Whiplash-Associated Disorders: redefining "whiplash" and its management. Spine (Phila Pa 1976). 1995;20:1S-73S.
2. Jansen GB, Edlund C, Grane P, Hildingsson C, Karlberg M, Link H, et al. Whiplash injuries: diagnosis and early management. The Swedish Society of Medicine and the Whiplash Commission Medical Task Force. Eur Spine J. 2008;17 Suppl 3:S355-417.
3. Sterling M. A proposed new classification system for whiplash associated disorders--implications for assessment and management. Man Ther. 2004;9:60-70.
4. Guzman J, Hurwitz EL, Carroll LJ, Haldeman S, Cote P, Carragee EJ, et al. A new conceptual model of neck pain: linking onset, course, and care: the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. J Manipulative Physiol Ther. 2009;32:S17-28.
5. Holm L, Cassidy JD, Sjogren Y, Nygren A. Impairment and work disability due to whiplash injury following traffic collisions. An analysis of insurance material from the Swedish Road Traffic Injury Commission. Scand J Public Health. 1999;27:116-23.
6. Hartling L, Brison RJ, Ardern C, Pickett W. Prognostic value of the Quebec Classification of Whiplash-Associated Disorders. Spine (Phila Pa 1976). 2001;26:36-41.
7. Carroll LJ, Holm LW, Hogg-Johnson S, Cote P, Cassidy JD, Haldeman S, et al. Course and prognostic factors for neck pain in whiplash-associated disorders (WAD): results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine (Phila Pa 1976). 2008;33:S83-92.
8. Curatolo M, Bogduk N, Ivancic PC, McLean SA, Siegmund GP, Winkelstein BA. The role of tissue damage in whiplash-associated disorders: discussion paper 1. Spine (Phila Pa 1976). 2011;36:S309-15.
9. Ferrari R, Russell AS, Carroll LJ, Cassidy JD. A re-examination of the whiplash associated disorders (WAD) as a systemic illness. Ann Rheum Dis. 2005;64:1337-42.
10. Berry H. Chronic whiplash syndrome as a functional disorder. Arch Neurol. 2000;57:592-4.
11. Barsky AJ, Borus JF. Functional somatic syndromes. Ann Intern Med. 1999;130:910-21.
12. Engel GL. The need for a new medical model: a challenge for biomedicine. Science. 1977;196:129-36.
13. Engel GL. The clinical application of the biopsychosocial model. Am J Psychiatry. 1980;137:535-44.
14. Schiltenwolf M, Beckmann NA. Whiplash disorder--is it a valid disease definition? Pain. 2013;154:2235.
15. Soltermann B. Studien des Schweizerischen Versicherungsverbandes SVV zum Thema Schleudertrauma. Schweizerische Ärztezeitung. 2004;49:2634-6.
16. Versteegen GJ, Kingma J, Meijler WJ, ten Duis HJ. Neck sprain after motor vehicle accidents in drivers and passengers. Eur Spine J. 2000;9:547-52.

17. Guez M, Hildingsson C, Nasic S, Toolanen G. Chronic low back pain in individuals with chronic neck pain of traumatic and non-traumatic origin: a population-based study. Acta Orthop. 2006;77:132-7.

18. Kamper SJ, Rebbeck TJ, Maher CG, McAuley JH, Sterling M. Course and prognostic factors of whiplash: a systematic review and meta-analysis. Pain. 2008;138:617-29.

19. Carroll LJ, Hogg-Johnson S, Cote P, van der Velde G, Holm LW, Carragee EJ, et al. Course and prognostic factors for neck pain in workers: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine (Phila Pa 1976). 2008;33:S93-100.

20. Holm LW, Carroll LJ, Cassidy JD, Skillgate E, Ahlbom A. Expectations for recovery important in the prognosis of whiplash injuries. PLoS Med. 2008;5:e105.

21. Scholten-Peeters GG, Verhagen AP, Bekkering GE, van der Windt DA, Barnsley L, Oostendorp RA, et al. Prognostic factors of whiplash-associated disorders: a systematic review of prospective cohort studies. Pain. 2003;104:303-22.

22. Carroll LJ, Ferrari R, Cassidy JD, Cote P. Coping and Recovery in Whiplash-associated Disorders: Early use of Passive Coping Strategies is Associated With Slower Recovery of Neck Pain and Pain-related Disability. Clin J Pain. 2014;30:1-8.

23. Kamper SJ, Hancock MJ, Maher CG. Optimal designs for prediction studies of whiplash. Spine (Phila Pa 1976). 2011;36:S268-74.

24. Scholz-Odermatt SM. Costs of whiplash claims in the years 2003-2012. Luzern: Central Office for Statistics in Accident Insurance (SSUV), 2014.

25. Chappuis G, Soltermann B. [Accident rates and Costs of Whiplash Associated Disorders. A Swiss pecularity?]. Rev Med Suisse. 2006;6:398-406.

26. Chappuis G, Soltermann B. Number and cost of claims linked to minor cervical trauma in Europe: results from the comparative study by CEA, AREDOC and CEREDOC. Eur Spine J. 2008;17:1350-7.

27. Ferrari R, Obelieniene D, Russell A, Darlington P, Gervais R, Green P. Laypersons' expectation of the sequelae of whiplash injury. A cross-cultural comparative study between Canada and Lithuania. Med Sci Monit. 2002;8:CR728-34.

28. Schrader H, Obelieniene D, Bovim G, Surkiene D, Mickeviciene D, Miseviciene I, et al. Natural evolution of late whiplash syndrome outside the medicolegal context. Lancet. 1996;347:1207-11.

29. Represas C, Vieira DN, Magalhaes T, Dias R, Frazao S, Suarez-Penaranda JM, et al. No cash no whiplash?: Influence of the legal system on the incidence of whiplash injury. J Forensic Leg Med. 2008;15:353-5.

30. Joslin CC, Khan SN, Bannister GC. Long-term disability after neck injury. a comparative study. J Bone Joint Surg Br. 2004;86:1032-4.

31. Sterling M, Hendrikz J, Kenardy J. Compensation claim lodgement and health outcome developmental trajectories following whiplash injury: A prospective study. Pain. 2010;150:22-8.

32. Sullivan MJ, Thibault P, Simmonds MJ, Milioto M, Cantin AP, Velly AM. Pain, perceived injustice and the persistence of post-traumatic stress symptoms during the course of rehabilitation for whiplash injuries. Pain. 2009;145:325-31.

33. Wenzel HG, Vasseljen O, Mykletun A, Nilsen TI. Pre-injury health-related factors in relation to self-reported whiplash: longitudinal data from the HUNT study, Norway. Eur Spine J. 2012;21:1528-35.

34. Mykletun A, Glozier N, Wenzel HG, Overland S, Harvey SB, Wessely S, et al. Reverse causality in the association between whiplash and symptoms of anxiety and depression: the HUNT study. Spine (Phila Pa 1976). 2011;36:1380-6.

35. Hurwitz EL, Carragee EJ, van der Velde G, Carroll LJ, Nordin M, Guzman J, et al. Treatment of neck pain: noninvasive interventions: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine (Phila Pa 1976). 2008;33:S123-52.

36. Holm LW, Carroll LJ, Cassidy JD, Hogg-Johnson S, Cote P, Guzman J, et al. The burden and determinants of neck pain in whiplash-associated disorders after traffic collisions: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine (Phila Pa 1976). 2008;33:S52-9.

37. Angst F, Francoise G, Verra M, Lehmann S, Jenni W, Aeschlimann A. Interdisciplinary rehabilitation after whiplash injury: an observational prospective outcome study. J Rehabil Med. 2010;42:350-6.

38. Robinson JP, Theodore BR, Dansie EJ, Wilson HD, Turk DC. The role of fear of movement in subacute whiplash-associated disorders grades I and II. Pain. 2013;154:393-401.

39. Cassidy JD, Carroll LJ, Cote P, Frank J. Does multidisciplinary rehabilitation benefit whiplash recovery?: results of a population-based incidence cohort study. Spine (Phila Pa 1976). 2007;32:126-31.

40. Cote P, Soklaridis S. Does early management of whiplash-associated disorders assist or impede recovery? Spine (Phila Pa 1976). 2011;36:S275-9.

41. Hartvigsen J. Musculoskeletal disorders and work disability. Pain. 2013;154:1904-5.

42. Gross DP, Russell AS, Ferrari R, Battie MC, Schopflocher D, Hu R, et al. Evaluation of a Canadian back pain mass media campaign. Spine (Phila Pa 1976). 2010;35:906-13.

43. Buchbinder R, Gross DP, Werner EL, Hayden JA. Understanding the characteristics of effective mass media campaigns for back pain and methodological challenges in evaluating their effects. Spine (Phila Pa 1976). 2008;33:74-80.

44. World Health Organisation. ICF - International Classification of Functioning, Disability and Health. ed. Geneva: World Health Organisation, 2001.

45. Waddell G. A new clinical model of low back pain and disability. In Waddell G ed. The back pain revolution. London: Churchill Livingstone, 1998:223-40.

46. Mechanic D. The concept of illness behavior. J Chronic Dis. 1962;15:189-94.

47. OECD. Sickness, Disability and Work: Breaking the Barriers. A Synthesis of findings across OECD countries. OECD Publishing. Paris, 2010.

48. Rainville J, Pransky G, Indahl A, Mayer EK. The physician as disability advisor for patients with musculoskeletal complaints. Spine (Phila Pa 1976). 2005;30:2579-84.

49. Oliveri M, Kopp G, Stutz K, Klipstein A, Zollikofer J. Basics of medical assessment of work capacity. Part 1 and 2. Schweiz Med Forum. 2006;6:420-31 (Part 1) and 48-54 (Part 2).

50. OECD. Sickness, Disability and Work: Breaking the Barriers. Norway, Poland and Switzerland. OECD Publishing Paris, 2006.

51. Raising expectations and increasing support: reforming welfare for future. The Stationary Office (TSO). Norwich (UK): Department for Work and Pensions, 2008.

52. ILO. Technical and ethical guidelines for workers' health surveillance. Occupational Safety and Health Series No. 72. International Labor Office [International Labor Organisation]. Geneva, 1998.

53. World Health Organisation. The Burden of Musculoskeletal Diseases at the Start of the New Millenium. Report of a WHO Scientific Group. TRS 919ed. Geneva, Switzerland: WHO, 2003.

54. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev. 1977;84:191-215.

55. Reneman MF, Jorritsma W, Schellekens JM, Goeken LN. Concurrent validity of questionnaire and performance-based disability measurements in patients with chronic nonspecific low back pain. J Occup Rehabil. 2002;12:119-29.

56. Holden G. The relationship of self-efficacy appraisals to subsequent health related outcomes: a meta-analysis. Soc Work Health Care. 1991;16:53-93.

57. van Abbema R, Lakke SE, Reneman MF, van der Schans CP, van Haastert CJ, Geertzen JH, et al. Factors associated with functional capacity test results in patients with non-specific chronic low back pain: a systematic review. J Occup Rehabil. 2011;21:455-73.

58. Rahman A, Reed E, Underwood M, Shipley ME, Omar RZ. Factors affecting self-efficacy and pain intensity in patients with chronic musculoskeletal pain seen in a specialist rheumatology pain clinic. Rheumatology (Oxford). 2008;47:1803-8.

59. Gonzalez-Calvo J, Gonzalez VM, Lorig K. Cultural diversity issues in the development of valid and reliable measures of health status. Arthritis Care Res. 1997;10:448-56.

60. Sloots M. Drop-out from rehabilitation in non-native patients with chronic non-specific low back pain. PhD thesis. Amsterdam: Vrije Universiteit Amsterdam, 2010.

61. Matheson LN. History, design characteristics, and uses of the pictorial activity and task sorts. J Occup Rehabil. 2004;14:175-95.

62. Matheson LN, Matheson M. Spinal Function Sort. Rating of Perceived Capacity. Test Booklet and Examiners Manualed. Trabuco Canyon, California: Performance and Capacity Testing, 1989.

63. U.S. Department of Labor. The Revised Handbook for Analyzing Jobs. 4th ed. Indianapolis: JIST Works, inc., 1991.

64. Sufka A, Hauger B, Trenary M, Bishop B, Hagen A, Lozon R, et al. Centralization of low back pain and perceived functional outcome. J Orthop Sports Phys Ther. 1998;27:205-12.

65. Robinson RC, Kishino N, Matheson L, Woods S, Hoffman K, Unterberg J, et al. Improvement in postoperative and nonoperative spinal patients on a self-report measure of disability: the Spinal Function Sort (SFS). J Occup Rehabil. 2003;13:107-13.

66. Kool JP, Oesch PR, Bachmann S, Knuesel O, Dierkes JG, Russo M, et al. Increasing days at work using function-centered rehabilitation in nonacute nonspecific low back pain: a randomized controlled trial. Arch Phys Med Rehabil. 2005;86:857-64.

67. Innes E, Hardwick M. Actual versus perceived lifting ability in healthy young men (18-25 years). Work. 2010;36:157-66.

68. Henchoz Y, de Goumoens P, So AK, Paillex R. Functional multidisciplinary rehabilitation versus outpatient physiotherapy for non specific low back pain: randomized controlled trial. Swiss Med Wkly. 2010;140:131-3.

69. Oliveri M. Functional Capacity Evaluation. In Gobelet C, Franchignoni F eds. Vocational Rehabilitation. Paris: Springer, 2005.

70. Gibson L, Strong J. The reliability and validity of a measure of perceived functional capacity for work in chronic back pain. J Occup Rehabil. 1996;6:159-75.

71. Matheson LN, Matheson ML, Grant J. Development of a measure of perceived functional ability. Journal of Occupational Rehabilitation. 1993;3:15-30.

72. Oesch PR, Hilfiker R, Kool JP, Bachmann S, Hagen KB. Perceived functional ability assessed with the spinal function sort: is it valid for European rehabilitation settings in patients with non-specific non-acute low back pain? Eur Spine J. 2010;19:1527-33.

73. Hoozemans MJ, van der Beek AJ, Frings-Dresena MH, van der Molen HF. Evaluation of methods to assess push/pull forces in a construction task. Appl Ergon. 2001;32:509-16.

74. Hansson GA, Balogh I, Bystrom JU, Ohlsson K, Nordander C, Asterland P, et al. Questionnaire versus direct technical measurements in assessing postures and movements of the head, upper back, arms and hands. Scand J Work Environ Health. 2001;27:30-40.

75. Balogh I, Orbaek P, Ohlsson K, Nordander C, Unge J, Winkel J, et al. Self-assessed and directly measured occupational physical activities--influence of musculoskeletal complaints, age and gender. Appl Ergon. 2004;35:49-56.

76. Unge J, Hansson GA, Ohlsson K, Nordander C, Axmon A, Winkel J, et al. Validity of self-assessed reports of occurrence and duration of occupational tasks. Ergonomics. 2005;48:12-24.

77. Ferrari R. The Whiplash Encyclopedia: The Facts and Myths of Whiplash. 2nd ed. Sudbury, MA: Jones and Bartlett Publisher, 2006.

78. Don AS, Carragee EJ. Is the self-reported history accurate in patients with persistent axial pain after a motor vehicle accident? Spine J. 2009;9:4-12.

79. Gatchel RJ. Psychosocial factors that can influence the self-assessment of function. J Occup Rehabil. 2004;14:197-206.

80. Smeets RJ, van Geel AC, Kester AD, Knottnerus JA. Physical capacity tasks in chronic low back pain: what is the contributing role of cardiovascular capacity, pain and psychological factors? Disabil Rehabil. 2007;29:577-86.

81. Huijnen IP, Verbunt JA, Peters ML, Delespaul P, Kindermans HP, Roelofs J, et al. Do depression and pain intensity interfere with physical activity in daily life in patients with Chronic Low Back Pain? Pain. 2010;150:161-6.

82. Isernhagen SJ. Functional capacity evaluation: rational, procedure, utility of the kinesiophysical approach. J Occup Rehabil. 1992;2:157-68.

83. Isernhagen SJ. Introduction to Functional Capacity Evaluation. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations. Chicago IL: American Medical Association, 2009:1-18.

84. Henchoz Y, de Goumoens P, Norberg M, Paillex R, So AK. Role of physical exercise in low back pain rehabilitation: a randomized controlled trial of a three-month exercise program in patients who have completed multidisciplinary rehabilitation. Spine (Phila Pa 1976). 2010;35:1192-9.

85. Oesch PR, Kool JP, Bachmann S, Devereux J. The influence of a Functional Capacity Evaluation on fitness for work certificates in patients with non-specific chronic low back pain. Work. 2006;26:259-71.

86. Wind H, Gouttebarge V, Kuijer PP, Sluiter JK, Frings-Dresen MH. Effect of Functional Capacity Evaluation information on the judgment of physicians about physical work ability in the context of disability claims. Int Arch Occup Environ Health. 2009;82:1087-96.

87. Soer R, Gerrits EH, Reneman MF. Test-retest reliability of a WRULD functional capacity evaluation in healthy adults. Work. 2006;26:273-80.

88. Harding VR, Williams AC, Richardson PH, Nicholas MK, Jackson JL, Richardson IH, et al. The development of a battery of measures for assessing physical functioning of chronic pain patients. Pain. 1994;58:367-75.

89. Golding LA. YMCA Fitness Testing and Assessment Manual. 4th ed. Champaign, IL: Human Kinetics, 2000.

90. Suva. Suva: an overview [Swiss Accident Insurance Fund] 2013. Available from: http://www.suva.ch/english/startseite-en-suva/suva-en-suva/ueberblick-en-suva.htm. Accessed 17.09.2013.

91. Robinson ME, Dannecker EA. Critical issues in the use of muscle testing for the determination of sincerity of effort. Clin J Pain. 2004;20:392-8.

92. Fishbain DA, Cutler R, Rosomoff HL, Rosomoff RS. Chronic pain disability exaggeration/malingering and submaximal effort research. Clin J Pain. 1999;15:244-74.

93. Sindhu BS, King PM. Assessing evaluee effort. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations: American Medical Association, 2009:195-226.

94. Denier-Bont F, Fischer V, Oesch P, Oliveri M. [Functional Capacity Evaluation: Course manual] ed. Bellikon: Verein IG Ergonomie, Swiss Association of Rehabilitation, 2007.

95. Innes E, Straker L. Validity of work-related assessments. Work. 1999;13:125-52.

96. King PM, Tuckwell N, Barrett TE. A critical review of functional capacity evaluations. Phys Ther. 1998;78:852-66.

97. Innes E, Straker L. Reliability of work-related assessments. Work. 1999;13:107-24.

98. Oesch P. Work-related evaluation and rehabilitation of patients with non-acute nonspecific low back pain. PhD thesis. Oslo, Norway: University of Oslo, 2012.

99. Gross D. Measurement properties of a functional capacity evaluation administered on workers' compensation claimants with low back pain. PhD thesis. Faculty of Rehabilitation Medicine. Edmonton, Canada: University of Alberta, 2003.

100. Gouttebarge V. Quality of functional capacity evaluation tests: a clinimetric approach. PhD thesis. Amsterdam, The Netherlands: Universiteit van Amsterdam, 2008.

101. Soer R. Functional capacity evaluation. Measurement qualities and normative values. PhD thesis. Groningen, The Netherlands: Rijksuniversiteit Groningen, 2009.

102. Hodselmans PA. Psychophysical capacity in non-specific chronic low back pain. PhD thesis. Groningen, The Netherlands: Rijkuniversiteit Groningen, 2009.

103. Reneman MF. Functional capacity evaluation in patients with chronic low back pain. Reliability and validity. PhD thesis. Groningen, The Netherlands: Rijksuniversiteit Groningen, 2004.

104. Lakke S. Work capacity of patients with chronic musculoskeletal pain. PhD thesis. Groningen, The Netherlands: Rijksuniversiteit Groningen, 2014.

105. Schiphorst Preuper HR. Determinants of disability and functional capacity in patients with chronic low back pain. PhD thesis. Groningen, The Netherlands: Rijksuniversiteit Groningen, 2011.

106. van Ittersum MW. Chronic musculoskeletal disorders: assessment and intervention. PhD thesis. Groningen, The Netherlands: Rijksuniversiteit Groningen, 2010.

107. Smeets RJEM. Active rehabilitation for chronic low back pain: cognitive-behavioral, physical or both? PhD thesis. Maastricht, The Netherlands: Maastricht University, 2006.

108. Innes E. Reliability and validity of functional capacity evaluations: an update. International Journal of Disability Management Research. 2006;1:135-48.

109. Gouttebarge V, Wind H, Kuijer PP, Frings-Dresen MH. Reliability and validity of Functional Capacity Evaluation methods: a systematic review with reference to Blankenship system, Ergos work simulator, Ergo-Kit and Isernhagen work system. Int Arch Occup Environ Health. 2004;77:527-37.

110. Lechner DE, Jackson JR, Roth DL, Straaton KV. Reliability and validity of a newly developed test of physical work performance. J Occup Med. 1994;36:997-1004.

111. Brouwer S, Dijkstra PU, Stewart RE, Goeken LN, Groothoff JW, Geertzen JH. Comparing self-report, clinical examination and functional testing in the assessment of work-related limitations in patients with chronic low back pain. Disabil Rehabil. 2005;27:999-1005.

112. Gross DP, Battie MC. Construct validity of a kinesiophysical functional capacity evaluation administered within a worker's compensation environment. J Occup Rehabil. 2003;13:287-95.

113. Gibson L, Strong J, Wallace A. Functional capacity evaluation as a performance measure: evidence for a new approach for clients with chronic back pain. Clin J Pain. 2005;21:207-15.

114. Lakke SE, Soer R, Geertzen JH, Wittink H, Douma RK, van der Schans CP, et al. Construct validity of functional capacity tests in healthy workers. BMC Musculoskelet Disord. 2013;14:180.

115. Soer R, van der Schans CP, Geertzen JH, Groothoff JW, Brouwer S, Dijkstra PU, et al. Normative values for a functional capacity evaluation. Arch Phys Med Rehabil. 2009;90:1785-94.

116. Bieleman HJ, van Ittersum MW, Groothoff JW, Oostveen JC, Oosterveld FG, van der Schans CP, et al. Functional capacity of people with early osteoarthritis: a comparison between subjects from the cohort hip and cohort knee (CHECK) and healthy ageing workers. Int Arch Occup Environ Health. 2010;83:913-21.

117. van Ittersum MW, Bieleman HJ, Reneman MF, Oosterveld FG, Groothoff JW, van der Schans CP. Functional capacity evaluation in subjects with early osteoarthritis of hip and/or knee; is two-day testing needed? J Occup Rehabil. 2009;19:238-44.

118. Reneman M, Wittink H, Gross DP. The scientific status of functional capacity evaluation. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations: American Medical Association, 2009:393-420.

119. Brouwer S, Reneman MF, Dijkstra PU, Groothoff JW, Schellekens JM, Goeken LN. Test-retest reliability of the Isernhagen Work Systems Functional Capacity Evaluation in patients with chronic low back pain. J Occup Rehabil. 2003;13:207-18.

120. Gouttebarge V, Wind H, Kuijer PP, Sluiter JK, Frings-Dresen MH. Reliability and agreement of 5 Ergo-Kit functional capacity evaluation lifting tests in subjects with low back pain. Arch Phys Med Rehabil. 2006;87:1365-70.

121. Soer R, van der Schans CP, Groothoff JW, Geertzen JH, Reneman MF. Towards consensus in operational definitions in functional capacity evaluation: a Delphi Survey. J Occup Rehabil. 2008;18:389-400.

Chapter 2

Which instruments can detect submaximal physical and functional capacity in patients with chronic nonspecific back pain? A systematic review

Suzan van der Meer
Maurizio A. Trippolini
Job van der Palen
Jan Verhoeven
Michiel F. Reneman

Spine. 2013; 38(25):E1608-15.

ABSTRACT

Study design: Systematic review

Objective: To evaluate the validity of instruments that claim to detect submaximal capacity when maximal capacity is requested in patients with chronic nonspecific musculoskeletal pain.

Summary of background data: Several instruments have been developed to measure capacity in patients with chronic pain. The detection of submaximal capacity can have major implications for patients. The validity of these instruments has never been systematically reviewed.

Methods: A systematic literature search was performed including the following databases: Web of Knowledge (including PubMed and Cinahl), Scopus and Cochrane. Two reviewers independently selected the articles based on the title and abstract according to the study selection criteria. Studies were included when they contained original data and when they objectified submaximal physical or functional capacity when maximal physical or functional capacity was requested. Two authors independently extracted data and rated the quality of the articles. The included studies were scored according to the subscales "criterion validity" and "hypothesis testing" of the COSMIN checklist. A Best Evidence Synthesis was performed.

Results: Seven studies were included, five of which used a reference standard for submaximal capacity. Three studies were of good methodological quality and validly detected submaximal capacity with specificity rates between 75% and 100%.

Conclusions: There is strong evidence that submaximal capacity can be detected in patients with chronic low back pain with a lumbar motion monitor or visual observations accompanying a Functional Capacity Evaluation lifting test.

INTRODUCTION

Detecting submaximal capacity when a maximal capacity is requested is challenging in patients with chronic musculoskeletal pain. Detection rates between 1% and 20% are reported, especially in the medico-legal context [1,2]. Instruments used to detect submaximal capacity, guide decisions that may have far-reaching implications in medical management but also for injury compensation claims. Therefore, it is of great importance to validly diagnose submaximal effort. Studies have been published about instruments that claim ability to discriminate maximal from submaximal capacity in patients with chronic musculoskeletal pain, but to our knowledge, a methodologically rigorous review of these studies has not been published.

Capacity is defined as the highest probable level of functioning that a person may reach in a domain at a given moment in a standardized environment [3]. Submaximal capacity can be referred to as malingering, disability exaggeration, symptom magnification syndrome or insincerity of effort. The Diagnostic and Statistical Manual of Mental Disorders (DSM) defines malingering as intentional production of false or grossly exaggerated physical or psychological disability, motivated by external incentives such as avoiding military duty, avoiding work, obtaining financial benefits, evading criminal prosecution or obtaining medication [4]. Symptom magnification syndrome is a self-destructive, socially reinforced behavioral response pattern consisting of reports or displays of symptoms which function to control the life circumstances of the sufferer [5]. Submaximal effort is related to muscle strength tests but is physiologically different from maximal effort [6]. Sincerity of effort has been described as a person's conscious motivation to perform optimally during evaluation and treatment [7]. There may be several reasons for a patient to put forth submaximal capacity, one of which being an adaptive reaction to avoid (increase of) pain. In this review, however, no distinction is made between intentional and unintentional reasons for submaximal capacity. There is a lack of clear definitions as to what constitutes submaximal capacity. In the International Classification of Functioning, Disability and Health (ICF) physical capacity and functional capacity are described [8,9]. Our definition of submaximal capacity is inspired by ICF: less than a maximal level of functioning on the physical or activity level that a person may reach in a domain at a given moment in a standardized environment. In this paper, the term submaximal is intentionally used and not malingering, insincerity, etc., because the reasons for submaximal capacity are beyond the scope of this study.

Maximal capacity tests serve as a standard against which to compare other measures. They play a key role in the assessment of maximal aerobic capacity or functional work capacity [10]. Some people are limited by cardiopulmonary, musculoskeletal and neuromuscular impairments and complaints such as dyspnea and pain. In those populations these instruments may be of limited use [10].

The aim of this systematic review was to identify the ability of instruments designed to detect submaximal physical or functional capacity when maximal capacity is requested in patients with nonspecific chronic musculoskeletal pain.

MATERIALS AND METHODS

Data sources and searches

Relevant studies were obtained through a computerized search of Web of Knowledge (including Medline and Cinahl), Scopus and Cochrane Library. The search included articles through October 10 2012 and used the following words: malingering, exaggeration, magnification, effort, discrepancies, submaximal, chronic pain (low back pain, whiplash injuries, fibromyalgia, neck pain) and is presented for the various databases in Appendix 2.1.

Studies in adults with nonspecific musculoskeletal chronic pain were included when they were: 1) written in English, German or Dutch; 2) contained original data; 3) objectified submaximal physical or functional capacity when maximal physical or functional capacity was requested. Studies describing mixed samples (e.g. subjects with pain and healthy subjects) or mixed methods (e.g. capacity test and self-report) were only included if the data of interest could be isolated.

Study selection

Two authors independently selected studies based on the title and abstract. Of potentially eligible studies a full copy was obtained. These articles were assessed for inclusion by two authors. Disagreements were resolved by discussion and if disagreement continued, a third person acted as an adjudicator. Additional reference tracking was performed. We hand-searched the reference lists of other relevant articles and eligible studies.

Data extraction and quality assessment

We used the COSMIN method to systematically evaluate the methodological quality of the studies [11]. The quality of the evidence for each study was assessed by using the COSMIN checklist Box H (criterion validity) or Box F (hypothesis testing) [11]. Two reviewers (SvdM and MT) independently assessed the methodological quality of the included studies. The quality criteria of Box H were used to score studies with a reference standard, whereas Box F was used to evaluate studies without a reference standard [11].

Data synthesis and analysis

To determine the overall quality of the measurement properties of the instruments, we synthesized the different studies by combining their results. In light of the study question, we were interested in test specificity. With lower specificity patients performing at maximal capacity will be rated as negative, and consequently incorrectly diagnosed as submaximal performers (false negative). With a lower sensitivity, patients performing at submaximal capacity will be rated as positive, and consequently incorrectly diagnosed as maximal performers (false positive). The possible overall ratings for a measurement property were positive (+), indeterminate (+/-) or negative (-), accompanied by levels of evidence, as was proposed by the Cochrane Review Back Group [12,13]. (Table 2.1) In the overall conclusion, because of their use of reference standards, criterion validity studies were preferred over hypothesis testing studies.

Table 2.1 Best evidence synthesis

Level	Rating	Criteria
Strong	+++	Consistent findings in multiple studies of good methodological quality OR in one study of excellent quality
Moderate	++	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality
Limited	+	One study of fair methodological quality
Conflicting	+/-	Conflicting findings
Unknown	?	Only studies of poor methodological quality

RESULTS

Study selection

The search strategy identified 2558 eligible studies. After screening the titles and abstracts, 29 potentially relevant studies were included. Of one study no full-text version could be obtained [14]. Twenty-one studies were excluded after reading the full text (Appendix 2.2). Seven studies were included (Figure 2.1).

Study characteristics

Information about patient characteristics, setting, blinding and test instruments is presented in Table 2.2. Six out of the seven studies assessed patients with low back pain. From the

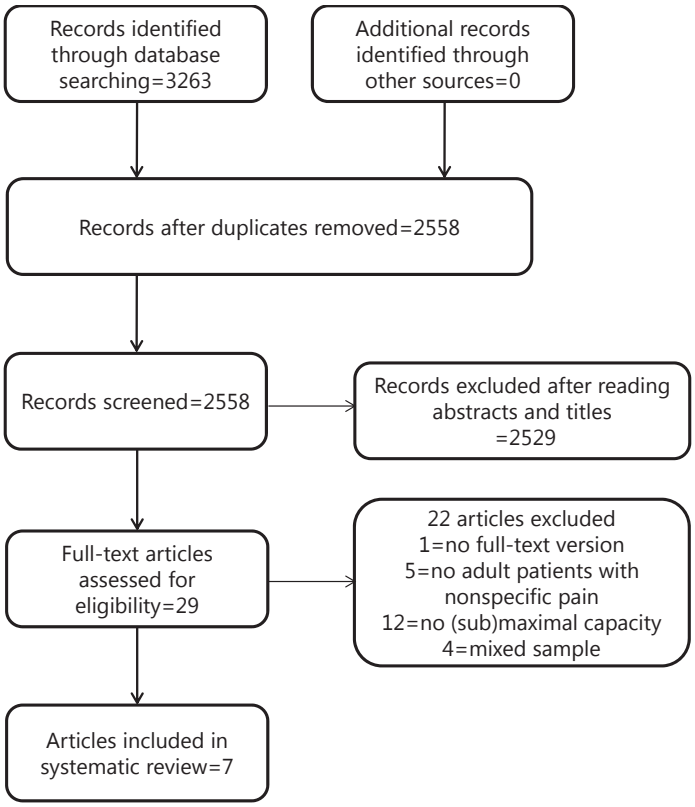


Figure 2.1 Flow diagram of study selection.

studies by Reneman et al. [15] and Dvir et al. [16], we included only the data which fulfilled the inclusion criteria. One of the review authors (MR) was also an author of one of the included trials. According to the Cochrane Review Guidelines and to avoid conflict of interest this author was not involved in the data analysis that involved his trial [12].

Instruments

Lemstra et al. [17] randomized 90 patients with low back pain in a 100% effort group and a 60% effort group. The patients performed a Functional Lumbar Lifting Test (PILE) and hand grip tests from a Functional Capacity Evaluation (FCE), in which 45 patients were asked to perform 60% effort on the tasks and 45 were asked to perform at 100% on the task. A blind tester gave an opinion as to whether the patient performed at 100% or 60% effort. This judgment was based on the analysis of all available data.

Reneman et al. [15] videotaped 16 patients with low back pain who performed a standardized lifting test as outlined in the Isernhagen Work System Functional Capacity (FCE). Sixty-three sets of lifting were edited on video and observed by nine trained observers who rated effort levels based on a rating scale.

Marras et al. [18] used a lumbar motion monitor to document the trunk motion characteristics of 100 patients with low back pain. The patients performed the test twice, one "sincere" trunk motion and one where they were asked to pretend that their pain was worse than it actually was. Judgment of submaximal effort was based on multivariate discriminant analyses and selected statistical models.

Dvir et al. [16] tested 25 patients with whiplash-related complaints using a cervical motion system for the rotation, lateral flexion, flexion and extension of the cervical column. The second time patients were asked to perform the tests whilst imagining that they were suffering from much more pain. Judgment of submaximal effort was done by the use of mixed effect models.

Luoto et al. [19] tested 23 patients with low back pain with a Lidoback isokinetic trunk dynamometer. The patients performed five trunk flexions at 100% effort, after three minutes rest they were asked to repeat the test at 50% of their maximal effort. The coefficient of variation was measured and differences between conditions tested with unpaired t-tests and Chi^2 tests.

Robinson et al. [20] performed an isometric lumbar extension task in 98 patients with chronic back pain and investigated the construct of symptom magnification with the results of Waddell signs, Minnesota Multiphasic Personality Inventory (MMPI) hysteria scale, MMPI hypochondriasis scale and the MMPI F-K index in a score. Judgment of submaximal capacity was done with the help of Pearson correlation coefficients.

In the study by Matheson et al. [21] 165 patients with low back pain underwent an FCE. An isometric grip strength measured with the JAMAR was performed and the examiner provided a score using the Symptom Magnification Rating.

Qualitative assessment

The results of the risk of bias assessment are presented in Tables 2.2 to 2.4. The blinding procedures were often not stated. In the studies by Lemstra et al. [17] and Reneman et al. [15], the observers were blinded. The studies with a reference standard were scored in box H (criterion validity) (Table 2.3). The studies of Robinson et al. [20] and Matheson et al. [21] were scored in Box F (hypothesis testing) because of their lack of a reference standard (Table 2.4). The reasons which led to the item scores are explained separately in the table. Lemstra et al. [17] asked their patients to perform maximal and also perform at 60% effort

Table 2.2 Study characteristics of the included studies

Study	Lemstra	Reneman	Marras	Dvir	Luoto	Robinson	Matheson
Diagnosis	Low back pain	Low back pain	Low back pain	Whiplash	Low back pain	Low back pain	Low back pain
Patients	90	16	100	25	23	98	165
Observations	1 measurements per patient	63 measurements in total	3 measurements in two ways per patient	6 measurements at two moments per patient	5 measurements in two ways per patient	7 measurement at two moments per patient	1 measurement per patient
Mean age ± SD	60% effort group 39.0 ± 10.4, 100% effort group 36.2 ± 12.7	39.6 ± 7.1	Men 38.4 ± 9.9 Women 37.4 ±11.2	37.1 ± 9.9	NR	40.3 ± 10.3	Males 38.3 ± 9.3 Females 41.9 ± 8.8
Male (%)	59	75	51	52	NR	74	66
Country	Canada	Netherlands	USA	Israel	Finland	USA	USA
Setting	Rehabilitation center	Rehabilitation center	Not stated	Not stated	Not stated	Rehabilitation center	Private rehabilitation center
Blinding therapists/ researchers	Yes/No	Yes/NR	No/NR	NR/NR	NR/NR	NR/NR	NR/NR
Test	Lumbar lifting test and hand grip tests from FCE with observators	FCE lift test with observators	Lumbar motion monitor and statistic models	Cervical motion system and SAS software	Back isokinetic trunk dynamometer	Lumbar extension isometric strength testing, Waddell signs and MMPI	FCE and Symptom Rating Scale
Key results	Sensitivity: 65.2% Specificity: 84.1%	Sensitivity: 6.9% Specificity: 100%	Sensitivity: 75% Specificity: 75%	Sub maximal capacity is hard to diagnose	Effort with a coefficient of variation (CV) 11-20% is hard to diagnose maximal or submaximal.	No strong support for the use of test-retest torque variability as a mean of detecting sub maximal effort	Grip strength consistency is not a significant predictor of symptom magnify-cation syndrome

NR = not reported, FCE = Functional Capacity Evalution.

Table 2.3 Results box H criterion validity

		Lemstra		Reneman		Marras		Dvir		Luoto	
1	Was the percentage of missing items given?	No missing items	E	No percentage, missing items were described	G	No missing items	E	Missing items not described	E	No missing items	E
2	Was there a description of how missing items were handled?	No missing items	E	They were not rated.	G	No missing items	E	No missing items mentioned	G	No missing items	E
3	Was the sample size included in the analysis adequate?	Good	G	Good	G	Adequate	E	Small	P	Poor	P
4	Can the criterion used or employed be considered as a reasonable "gold standard"?	Asked to give 60% effort compared to 100% effort within subjects	E	Sub maximal performance compared to better maximal performance of within subjects	E	Act that pain was worse than it actually was compared with maximum performance	G	Imagine that your pain is worse compared with maximum performance	G	Asked to give the second time 50% effort instead of 100%	E
5	Where there any important flaws in the design or methods of the study?	None	E	None	E	None	E	None	E	None	E
6	For continuous scores: Were correlations, or the area under the receiver operating curve calculated?	NA		NA		NA		NA		No	P
7	For dichotomous scores: Were sensitivity and specificity determined?	Yes	E	Yes	E	Yes	E	NA		NA	
	Quality score	**Good**	**G**	**Good**	**G**	**Good**	**G**	**Poor**	**P**	**Poor**	**P**

E=Excellent, G=Good, F=Fair, P=Poor, NA= not applicable.

Table 2.4 Box F results hypotheses testing

		Robinson		Matheson	
1	Was the percentage of missing items given?	No missing items	G	No missing items	E
2	Was there a description of how missing items were handled?	No missing items mentioned	G	No missing items	E
3	Was the total sample size included in the analysis adequate	Good	G	Adequate	E
4	Were hypotheses regarding correlation or mean differences formulated a priori?	Hypotheses at end of introduction	G	Hypotheses at end of paragraph 2	E
5	Was the expected direction of correlations or mean differences included in the hypotheses?	Symptom magnification positive related to torque variability	E	Less than full effort is indication of symptom magnification	E
6	Was the expected absolute or relative magnitude of correlations or mean differences included in the hypotheses?	Not stated	G	Not stated	G
7	Was an adequate description provided of the comparator instrument?	The measure of torque variability is described, but the amount of torque variability is not described.	P	Worse score compared to full-effort on performance level and repeatability	G
8	Were the measurement properties of the comparator instrument adequately described?	Adequate description in introduction and methods.	E	References in introduction	G
9	Were there any important flaws in the design of method of the study?	None	E	None	E
10	Were design and statistical methods adequate for the hypotheses to be tested?	Appropriate	E	Appropriate	E
	Quality score	**Poor**	**P**	**Good**	**G**

E=Excellent, G=Good, F=Fair, P=Poor.

and Luoto et al. [19] asked their patients also to perform at 50% effort. Reneman et al. [15] used observations of submaximal performance followed by higher performance. The studies by Marras et al. [18] and Dvir et al. [16] asked their patients to imagine that their pain was worse than it actually was.

Based on the scoring system of the COSMIN checklist Marras et al. [18], Lemstra et al. [17] and Reneman et al. [15] scored GOOD and Dvir et al. [16] and Luoto et al. [19] scored POOR. Matheson et al. [21] scored GOOD and Robinson et al. [20] scored POOR. Cohen's kappa for overall agreement between the two reviewers was 0.77, which is considered to represent substantial agreement. Full agreement for all criteria (k=1.0) was reached during the consensus meeting.

Data synthesis and analysis

Three studies dichotomized their tests and used a sensitivity and specificity analysis. Lemstra et al. [17] reported a sensitivity of 65% and a specificity of 84%, which means that the test will identify 65% of all patients performing at a maximal level (sensitivity), and that the test will identify 84% of all patients performing at a submaximal level (specificity). Reneman et al. [15] reported a sensitivity of 7% and a specificity of 100%, and mentioned that they were uncertain whether their patients performed maximally (because of the absence of a reference standard for maximal performance). Marras et al. [18] reported both a sensitivity and specificity of 75%. Consented cutoff values for acceptable specificity and sensitivity are not available: however, with lower specificity patients performing at maximal capacity will be rated as false negative, and consequently incorrectly diagnosed as submaximal performers. With a lower sensitivity, patients performing at submaximal capacity will be rated as false positive, and consequently incorrectly diagnosed as maximal performers. These three studies were rated positive. The study by Dvir et al. [16] concluded that there was a relatively small and stable compression of cervical motion when patients simulated pain, so with their instrument, submaximal capacity was hard to diagnose. Luoto et al. [19] concluded that effort with a coefficient of variation between 11–20% is hard to diagnose maximal or submaximal. Robinson et al. [20] concluded that there is no strong support for the use of test-retest torque variability as a means of detecting submaximal effort. Matheson et al. [21] claimed that grip strength consistency is not a significant predictor of symptom magnification syndrome. The ratings based on the best evidence synthesis are stated in Table 2.5. Finally, there is in the criterion validity strong evidence that submaximal capacity can be detected in patients with chronic low back pain with a FCE lifting test or a lumbar motion monitor and there is moderate evidence in the case of hypothesis testing that submaximal capacity cannot be detected in patients with chronic low back pain.

Table 2.5 Data synthesis of the included studies

Study	Box	Rating test instrument	Rating methodological quality
Lemstra	H	+	Good
Reneman	H	+	Good
Marras	H	+	Poor
Dvir	H	-	Good
Luoto	H	-	Poor
Robinson	F	-	Poor
Matheson	F	-	Good

DISCUSSION

Based on the results of three good quality studies there is strong evidence that submaximal capacity can be detected in patients with chronic low back pain with visual observations accompanying a FCE lifting test or a lumbar motion monitor.

In two studies with a reference standard and good methodological quality, visual observations accompanying FCE was used as the test instrument. The FCE is an instrument used to determine functional capacity [6,22]. FCEs are applied in rehabilitation, occupational and insurance medicine [23,24]. For further diagnostic studies on submaximal effort in patients with chronic musculoskeletal pain, the use of FCE including a physical effort determination by trained observers should be considered, over a method using statistical cut off values only. A reference standard could also be a lumbar motion monitor or another sophisticated testing device or procedure, for example superimposed electrical stimulation [25]. The instruments enquire training to use it in a correct way, but provide added clinical value. The specificity of the studies varied between 75% and 100%. False negative diagnoses can have major implications and it is debatable if a specificity of 75% is sufficient to justify its use. Also, there are several extraneous variables that may influence muscle testing [26]. Several factors such as an unfamiliar testing environment or testing apparatus fear of pain and/or (re)injury, anxiety, depression, anger, work satisfaction, self-reported disability, motivation, medication consumption, and pain have been reported to influence the maximum capacity [26]. Those factors should be considered, when diagnosing submaximal capacity.

When comparing the results of the current systematic review with the findings of Fishbain et al. [2], they used a broader definition of submaximal capacity and therefore included more articles. They concluded that isometric strength testing and the use of the coefficient of variation did not reliably discriminate between full and submaximal effort, but isokinetic

testing did, which is in contrast to our conclusion. In our review, however, the methodological quality of the study using isokinetic testing was rated as poor [19]. Because Fishbain et al. [2] did not perform a qualitative rating of the included studies, insufficiently designed and reported diagnostic studies may have influenced their results and conclusions. In healthy people, sincerity of effort was reviewed by Robinson et al. [26]. They stated that submaximal effort can be reliably discriminated from maximal effort in muscle testing with the help of statistical models. In general, submaximal effort conditions will reliably show greater variability than maximal effort conditions [26]. However, the clinical utility of variability cut-offs has still not been validated. Moreover, several studies have an inadequate sample size, unknown generalizability or other explanatory factors such as pain or fear of injury that should be considered in evaluating a person's sincerity of effort [26]. In neuropsychology, detection of submaximal effort has also received much attention [27,28]. However, it appears that an acceptable reference standard for methods that claim to detect submaximal capacity in neuropsychology has not yet been developed [29]. An example of a reference standard for submaximal functional capacity in our review is that if a person has lifted 10, 20, 30 and 40 kg within a five minutes session, then 10, 20 and 30 kg are submaximal efforts [15]. Hence, if patients are asked to perform submaximal and maximal, a reference standard for submaximal capacity is available.

This is the first systematic review about submaximal capacity in which definitions of submaximal physical and functional capacity were clearly described. This systematic review was performed following highly transparent procedures, using recommended checklists for the assessment of the methodological quality of health related outcome measures and by reporting a best evidence synthesis. In most of the included studies, there might have been some risk of bias, because procedures to "blind" researchers and testers were not described. Although we used clear definitions for submaximal physical and functional capacity, the authors of the included articles used their own terminology with regard to malingering, symptom magnification and effort. There is not yet a clear general definition of these terms. It is unknown to what extent either better blinding strategies or clear definitions would have affected the conclusions of this systematic review.

CONCLUSIONS

In conclusion, this systematic review has identified few instruments that validly detect submaximal capacity in clinical samples with chronic pain. Knowing the relevance for the individual and society to accurately differentiate submaximal from maximal capacity, some major advances should be made to perform methodologically well-designed diagnostic studies with large clinical samples and practical instruments.

REFERENCES

1. Greve KW, Ord JS, Bianchini KJ, et al. Prevalence of malingering in patients with chronic pain referred for psychologic evaluation in a medico-legal context. Arch Phys Med Rehabil. 2009;90(7):1117-26.
2. Fishbain DA, Cutler R, Rosomoff HL, Rosomoff RS. Chronic pain disability exaggeration/malingering and submaximal effort research. Clin J Pain. 1999;15(4):244-74.
3. Soer R, van der Schans CP, Groothoff JW, et al. Towards consensus in operational definitions in functional capacity evaluation: A delphi survey. J Occup Rehabil. 2008;18(4):389-400.
4. American Psychiatric Association, ed. Diagnostic and statistical manual of mental disorders. Washington DC: American Psychiatric Press; 2000.
5. Matheson L. Symptom magnification syndrome structured interview: Rationale and procedure. J Occup Rehabil. 1991;1(1):43-56.
6. Sindhu BS, King PM. Assessing evaluee effort. In: Guide to the evaluation of functional ability. how to request, interpret and apply functional capacity evaluations. United States of America: American Medical Association; 2009:195-226.
7. Lechner DE, Bradbury SF, Bradley LA. Detecting sincerity of effort: A summary of methods and approaches. Phys Ther. 1998;78(8):867-88.
8. Stucki G, Ewert T, Cieza A. Value and application of the ICF in rehabilitation medicine. Disabil Rehabil. 2002;24(17):932-8.
9. World Health Organization. International classification of functioning, disability and health (ICF). Geneva. 2001.
10. Noonan V, Dean E. Submaximal exercise testing: Clinical application and interpretation. Phys Ther. 2000;80(8):782-807.
11. Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. Quality of Life Research. 2012;21:651-7.
12. Furlan AD, Pennick V, Bombardier C, et al. Editorial Board Cochrane Back Revi. 2009 updated method guidelines for systematic reviews in the cochrane back review group. Spine. 2009;34(18):1929-41.
13. Van Tulder M, Furlan A, Bombardier C, et al. Updated method guidelines for systematic reviews in the cochrane collaboration back review group. Spine. 2003;28(12):1290-9.
14. Miller T, Allen G, Gandevia S. Muscle force, perceived effort, and voluntary activation of the elbow flexors assessed with sensitive twitch interpolation in fibromyalgia. J Rheumatol. 1996;23(9):1621-7.
15. Reneman MF, Fokkens AS, Dijkstra PU, et al. Testing lifting capacity: Validity of determining effort level by means of observation. Spine. 2005;30(2):E40-6.
16. Dvir Z, Gal-Eshel N, Shamir B, et al. Simulated pain and cervical motion in patients with chronic disorders of the cervical spine. Pain Research and Management. 2004;9(3):131-6.
17. Lemstra M, Olszynski W, Enright W. The sensitivity and specificity of functional capacity evaluations in determining maximal effort - A randomized trial. Spine. 2004;29(9):953-9.

18. Marras WS, Lewis KEK, Ferguson SA, et al. Impairment magnification during dynamic trunk motions. Spine. 2000;25(5):587-95.

19. Luoto S, Hupli M, Alaranta H, et al. Isokinetic performance capacity of trunk muscles. part II: Coefficient of variation in isokinetic measurement in maximal effort and in submaximal effort. Scand J Rehabil Med. 1996;28(4):207-10.

20. Robinson ME, O'Connor PD, MacMillan M, et al. Physical and psychosocial correlates of test-retest isometric torque variability in patients with chronic low back pain. J Occup Rehabil. 1992;2(1):11-8.

21. Matheson LN, Bohr PC, Hart DL. Use of maximum voluntary effort grip strength testing to identify symptom magnification syndrome in persons with low back pain. Journal of Back and Musculoskeletal Rehabilitation. 1998;10(3):125-35.

22. Gouttebarge V, Wind H, Kuijer PPFM, et al. Reliability and validity of functional capacity evaluation methods: A systematic review with reference to blankenship system, ergos work simulator, ergo-kit and isernhagen work system. Int Arch Occup Environ Health. 2004;77(8):527-37.

23. Genovese E, Galper JS. Guide to the evaluation of functional ability. how to request, interpret and apply functional capacity evaluations. United States of America: American Medical Association; 2009

24. Wind H, Gouttebarge V, Kuijer PPFM, Sluiter JK, et al. Effect of functional capacity evaluation information on the judgment of physicians about physical work ability in the context of disability claims. Int Arch Occup Environ Health. 2009;82(9):1087-96.

25. Verbunt JA, Seelen HA, Vlaeyen JW, et al. Pain-related factors contributing to muscle inhibition in patients with chronic low back pain: An experimental investigation based on superimposed electrical stimulation. Clin J Pain. 2005;21(3):232-40.

26. Robinson ME, Dannecker EA. Critical issues in the use of muscle testing for the determination of sincerity of effort. Clin J Pain. 2004;20(6):392-8.

27. Bianchini KJ, Greve KW, Glynn G. On the diagnosis of malingered pain-related disability: Lessons from cognitive malingering research. Spine Journal. 2005;5(4):404-17.

28. Bianchini K, Mathias C, Greve K. Symptom validity testing: A critical review. Clin Neuropsychol. 2001;15(1):19-45.

29. Heilbronner RL, Sweet JJ, Morgan JE, et al. Conference Participants. American academy of clinical neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. Clin Neuropsychol. 2009;23(7):1093-129.

Appendix 2.1 Search strategy

Database	Search terms	Include	Exclude
Web of Knowledge	chronic pain[MeSH] OR back pain [MeSH] OR neck pain [MeSH] OR whiplash injuries[MeSH] OR fibromyalgia [MeSH] (TOPIC) AND malingering[MeSH] OR exaggeration[tiab] OR magnification[tiab] OR effort[tiab] OR discrepancies[tiab] OR submaximal[tiab] (TOPIC)	1. articles 2. English, German, Dutch from languages	1. neuroscience and neurology
Scopus	(chronic pain OR back pain OR neck pain OR whiplash OR fibromyalgia) AND (malingering OR exaggeration OR magnification OR effort OR discrepancies OR submaximal) (TAK)	1. articles	
Cochrane	chronic pain OR back pain OR neck pain OR whiplash injuries OR fibromyalgia (TAK) AND malingering OR exaggeration OR magnification OR effort OR discrepancies OR submaximal (TAK)		

Appendix 2.2 Excluded studies

Author	Title	Country	Reason exclusion
Khalil [1]	Acceptable Maximum Effort (AME) - a psychophysical measure of strength in back pain patients.	U.S.A.	No adult patients with nonspecific musculoskeletal chronic pain
Duque [2]	Aerobic fitness and limiting factors of maximal performance in chronic low back pain patients	Colombia	No adult patients with nonspecific musculoskeletal chronic pain
Ng [3]	Functional roles of abdominal and back muscles during isometric axial rotation of the trunk.	Australia	No adult patients with nonspecific musculoskeletal chronic pain
Robinson [4]	Lumbar iEMG during isotonic exercise: Chronic low back pain patients versus controls.	U.S.A.	No adult patients with nonspecific musculoskeletal chronic pain
Akebi [5]	Factors affecting the variability of the torque curves at isokinetic trunk strength testing.	Japan	No study that objectified submaximal capacity when maximal capacity was requested
Dvir [6]	Trunk extension effort in patients with chronic low back dysfunction.	Australia	No study that objectified submaximal capacity when maximal capacity was requested
Hazard [7]	Disability exaggeration as a predictor of functional restoration outcomes for patients with chronic low-back pain.	Denmark	No study that objectified submaximal capacity when maximal capacity was requested
Kaplan [8]	Maximal effort during Functional Capacity Evaluations: An examination of psychological factors.	U.S.A.	No study that objectified submaximal capacity when maximal capacity was requested
Oesch [9]	Comparison of two methods for interpreting lifting performance during functional capacity evaluation.	Switzerland	No study that objectified submaximal capacity when maximal capacity was requested
Reid [10]	Isokinetic trunk-strength deficits in people with and without low-back pain: A comparative study with consideration of effort.	U.S.A.	No study that objectified submaximal capacity when maximal capacity was requested
Ylinen [11]	Association of neck pain, disability and neck pain during maximal effort with neck muscle strength and range of movement in women with chronic non-specific neck pain.	Finland	No study that objectified submaximal capacity when maximal capacity was requested

Appendix 2.2 continues on next page

Appendix 2.2 *Continued*

Author	Title	Country	Reason exclusion
Lindh [12]	Studies on maximal voluntary muscle-contraction in patients with fibromyalgia.	Sweden	No study that objectified submaximal capacity when maximal capacity was requested
Oddsson [13]	Activation imbalances in lumbar spine muscles in the presence of chronic low back pain.	U.S.A.	No study that objectified submaximal capacity when maximal capacity was requested
O'Leary [14]	A new method of isometric dynamometry for the craniocervical flexor muscles.	Australia	No study that objectified submaximal capacity when maximal capacity was requested
Roe [15]	Muscle activation during isometric contractions in workers with unilateral shoulder myalgia.	Norway	No study that objectified submaximal capacity when maximal capacity was requested
Newton [16]	Trunk strength testing with Iso-Machines: Part 2: Experimental evaluation of the Cybex II back testing system in normal subjects and patients with chronic low back pain.	Scotland	No study that objectified submaximal capacity when maximal capacity was requested
Da Silva [17]	Back muscle strength and fatigue in healthy and chronic low back pain subjects: A comparative study of 3 assessment protocols.	Canada	No study that objectified submaximal capacity when maximal capacity was requested
Schapmire [18]	Simultaneous bilateral hand strength testing in a client population, part I: Diagnostic, observational and subjective complaint correlates to consistency of effort.	U.S.A.	Contained mixed samples where data on the relevant subgroups could not be isolated
Ruan [19]	Functional Capacity Evaluations in persons with spinal disorders: Predicting poor outcomes on the Functional Assessment Screening Test (FAST).	U.S.A.	Contained mixed samples where data on the relevant subgroups could not be isolated
Hutten [20]	Differences in treatment outcome between subgroups of patients with chronic low back pain using lumbar dynamometry and psychological aspects.	Netherlands	Contained mixed samples where data on the relevant subgroups could not be isolated
Hutten [21]	Distribution of psychological aspects in subgroups of chronic low back pain patients divided on the score of physical performance.	Netherlands	Contained mixed samples where data on the relevant subgroups could not be isolated

References Appendix 2.2

1. Khalil TM, Goldberg ML, Asfour SS, et al. Acceptable maximum effort (ame) - a psychophysical measure of strength in back pain patients. Spine. 1987;12(4):372-376.
2. Leonardo Duque I, Parra J, Duvallet A. Aerobic fitness and limiting factors of maximal performance in chronic low back pain patients. Journal of Back and Musculoskeletal Rehabilitation. 2009;22(2):113-119.
3. Ng JKF, Parnianpour M, Richardson CA, et al. Functional roles of abdominal and back muscles during isometric axial rotation of the trunk. Journal of Orthopaedic Research. 2001;19(3):463-471.
4. Robinson ME, Cassisi JE, O'Connor PD, et al. Lumbar iEMG during isotonic exercise: Chronic low back pain patients versus controls. J Spinal Disord. 1992;5(1):8-15.
5. Akebi T, Saeki S, Hieda H, et al. Factors affecting the variability of the torque curves at isokinetic trunk strength testing. Arch Phys Med Rehabil. 1998;79(1):33-35.
6. Dvir Z, Keating JL. Trunk extension effort in patients with chronic low back dysfunction. Spine. 2003;28(7):685-692.
7. Hazard RG, Bendix A, Fenwick JW. Disability exaggeration as a predictor of functional restoration outcomes for patients with chronic low-back pain. Spine. 1991;16(9):1062-1067.
8. Kaplan G, Wurtele S, Gillis D. Maximal effort during functional capacity evaluations: An examination of psychological factors. Arch Phys Med Rehabil. 1996;77(2):161-164.
9. Oesch P, Meyer K, Bachmann S, et al. Comparison of two methods for interpreting lifting performance during functional capacity evaluation. Phys Ther. 2012;92(9):1130-1140.
10. Reid S, Hazard RG, Fenwick JW. Isokinetic trunk-strength deficits in people with and without low-back pain: A comparative study with consideration of effort. J Spinal Disord. 1991;4(1):68-72.
11. Ylinen J, Takala E-, Kautiainen H, et al. Association of neck pain, disability and neck pain during maximal effort with neck muscle strength and range of movement in women with chronic non-specific neck pain. European Journal of Pain. 2004;8(5):473-478.
12. Lindh M, Johansson L, Hedberg M, Grimby G. Studies on maximal voluntary muscle-contraction in patients with fibromyalgia. Arch Phys Med Rehabil. 1994;75(11):1217-1222.
13. Oddsson LIE, De Luca CJ. Activation imbalances in lumbar spine muscles in the presence of chronic low back pain. J Appl Physiol. 2003;94(4):1410-1420.
14. O'Leary S, Vicenzino B, Jull G. A new method of isometric dynamometry for the craniocervical flexor muscles. Phys Ther. 2005;85(6):556-564.
15. Røe C, Knardahl S, Vøllestad NK. Muscle activation during isometric contractions in workers with unilateral shoulder myalgia. J Musculoskeletal Pain. 2000;8(4):57-73.
16. Newton M, Thow M, Somerville D, et al. Trunk strength testing with iso-machines: Part 2: Experimental evaluation of the cybex II back testing system in normal subjects and patients with chronic low back pain. Spine. 1993;18(7):812-824.
17. Da Silva Jr. RA, Arsenault AB, Gravel D, et al. Back muscle strength and fatigue in healthy and chronic low back pain subjects: A comparative study of 3 assessment protocols. Arch Phys Med Rehabil. 2005;86(4):722-729.
18. Schapmire DW, St James JD, Feeler L, et al. Simultaneous bilateral hand strength testing in a client population, part I: Diagnostic, observational and subjective complaint correlates to consistency of effort. Work-a Journal of Prevention Assessment & Rehabilitation. 2010;37(3):309-320.
19. Ruan CM, Haig AJ, Geisser ME, et al. Functional capacity evaluations in persons with spinal disorders: Predicting poor outcomes on the functional assessment screening test (FAST). J Occup Rehabil. 2001;11(2):119-132.
20. Hutten MMR, Hermens HJ, Zilvold G. Differences in treatment outcome between subgroups of patients with chronic low back pain using lumbar dynamometry and psychological aspects. Clin Rehabil. 2001;15(5):479-488.
21. Hutten MMR, Hermens HJ, Ijzerman MJ, et al. Distribution of psychological aspects in subgroups of chronic low back pain patients divided on the score of physical performance. International Journal of Rehabilitation Research. 1999;22(4):261-268.

Chapter 3

Reliability of clinician rated physical effort determination during functional capacity evaluation in patients with chronic musculoskeletal pain

Maurizio A. Trippolini
Pieter U. Dijkstra
Beatrice Jansen
Peter Oesch
Jan H. B. Geertzen
Michiel F. Reneman

Journal of Occupational Rehabilitation. 2014; 24: 361-369.

ABSTRACT

Introduction: Functional Capacity Evaluation (FCE) can be used to make clinical decisions regarding fitness-for-work. During FCE the evaluator attempts to assess the amount of physical effort of the patient. The aim of this study is to analyze the reliability of physical effort determination using observational criteria during FCE.

Methods: Twenty-one raters assessed physical effort in 18 video-recorded FCE tests independently on two occasions, 10 months apart. Physical effort was rated on a categorical four-point physical effort determination scale (P_{ED}) based on the Isernhagen criteria, and a dichotomous submaximal effort determination scale (S_{ED}). Cohen's Kappa, squared weighted Kappa and % agreement were calculated.

Results: Kappa values for intra-rater reliability of P_{ED} and S_{ED} for all FCE tests were 0.49 and 0.68 respectively. Kappa values for inter-rater reliability of P_{ED} for all FCE tests in the first and the second session were 0.51, and 0.72, and for S_{ED} Kappa values were 0.68 and 0.77 respectively. The inter-rater reliability of P_{ED} ranged from κ=0.02 to κ=0.99 between FCE tests. Acceptable reliability scores (κ>0.60, agreement ≥80%) for each FCE test were observed in 38% of scores for P_{ED} and 67% for S_{ED}. On average material handling tests had a higher reliability than postural tolerance and ambulatory tests.

Conclusion: Dichotomous ratings of submaximal effort are more reliable than categorical criteria to determine physical effort in FCE tests. Regular education and training may improve the reliability of observational criteria for effort determination.

INTRODUCTION

Individuals suffering from chronic nonspecific musculoskeletal pain (CMP) such as back and neck pain are often restricted in performing activities of daily living and work [1,2]. The financial burden of CMP on society arises mainly due to indirect costs because of temporary or permanent work disability. Work disability due to CMP may be associated with reduced activity levels and work performance [3,4]. Functional capacity evaluation (FCE) in addition to self-reported measures have been recommended for a comprehensive assessment of physical work performance for persons with CMP [5-8].

FCE employs physical performance tests such as lifting, postural tolerance tests, repetitive movements, and ambulation to assess work-related functioning [9]. Discrepancies in FCE outcomes and the physical workload of a patient may be addressed in rehabilitation to restore this imbalance [10-12]. Moreover, FCEs are used to evaluate the effects of rehabilitation and determine fitness-for-work, and as such FCEs may facilitate the return-to-work process or prelude case closure [13-17].

To determine physical capacity during the FCE the patient must perform to his or her maximum level of physical ability. The level of physical effort during FCE is estimated by the evaluator, based on observational criteria during material and non-material handling tests [9,18]. Submaximal effort is assumed when a person stops a FCE test before the criteria indicative of maximal effort are observed. Because clinical decision-making is based on the results of FCE, sound clinimetric properties of observational criteria are required to determine physical effort. Acceptable reliability of physical effort determination FCE tests such as lifting has been reported [19,20]. However, the reliability of non-material handling tests such as kneeling and forward bending has rarely been studied [21-25]. Moreover, most studies on lifting tests were performed by FCE experts, which limits the generalizability and applicability of the study results among less experienced raters [25-27].

The aim of this study was to determine the intra- and inter-rater reliability of physical effort determination of FCE tests in patients with CMP. A second aim was to investigate whether an increase in rater experience would alter the reliability of physical effort determination.

METHODS

Procedures, patients and video sequences

Video tape-recordings were taken during FCEs, performed in a work rehabilitation setting. FCE tests were performed according to the Isernhagen test procedure, which claims to measure

a person's physical capacity to safely engage in work-related activity [28]. Four patients (3 with non-specific low back pain and 1 with non-specific neck pain, mean age 35.5 years, range 21 to 49 years) were recruited based on convenience. All patients were instructed how to perform the test, and that they were expected to perform maximally. Testing could be terminated for four reasons: the participant stopped because of, for example, pain; the observer deemed testing to have become over safe maximum based on criteria for effort determination (Appendices 3.1, 3.2); heart rate exceeded 85% of the age-related maximum (220 minus age of participant); or a predefined time limit was reached. All patients gave written consent to be video-recorded. Eighteen videos from 11 FCE tests with a total duration of 28 minutes were selected. The videos were mute recorded. For each test information was provided on a standardized form regarding heart rate at the beginning and end of the test, and weight lifted in kilograms (for material handling tests) or duration (for static posture, or walking, stair climbing).

Raters

A convenience sample of 21 physiotherapists (11 female, 10 male) from Bellikon rehabilitation clinic (Switzerland) served as a representative sample of raters. Nineteen had attended the official 2-day FCE training course provided by the Swiss Rehabilitation Association [18]. Prior to the study all had performed at least ten 1-day FCEs in the previous year (median 30, interquartile range (IQR): 20 to 33) and had a minimum of 1 year work experience in work rehabilitation (median 3, IQR: 2 to 3), and a minimum professional practice experience of 1 year (median 5 years, IQR: 3 to 12.5).

Physical effort determination during FCE tests

The 18 videos were shown in a classroom to all the raters at the same time. Prior to the showing the raters were instructed about the procedure of the rating. The ratings of physical effort were filled in a standardized form with a pencil. The videos consisted of 18 tests. When a test was finished and all participants had rated that test, then the next test was shown. Raters were not allowed rewind the video or to stop a video while a test was shown. Each video was shown once per session. Raters were blinded each other's ratings. Each video was rated according to observational criteria indicative of physical effort for material handling tests as "light to moderate", "heavy" or "maximal" (Appendix 3.1). Observational criteria for postural tolerance tests and ambulation tests were rated on a scale from "No or slight functional problem/limitation", "some functional problem/limitation" to "substantial functional problem/limitation" (Appendix 3.2). This categorical scale was termed physical effort determination (P_{ED}) scale. If a test was performed unsafely it was classified as "over safe maximum", when

observed performance exceeded the maximum observational criteria for physical effort level during work-related tasks (Appendices 3.1+3.2). Tests were scored as "not classifiable" when the patient interrupted the FCE test at the very start or the observed effort was not clearly interpretable to the raters and no conclusions could be drawn. Submaximal effort was assumed when a patient stopped a material or non-material handling test before the FCE rater observed sufficient criteria indicative of maximal weight, or significant functional problems/limitation as described in Appendices 3.1+3.2. This dichotomous scale was termed submaximal effort determination (S_{ED})

Maximal effort was defined as the highest safe ability of a person during a FCE test [9]. An FCE was considered safe when no formal complaints of injury or serious adverse effects were filed by the patients, and when increased symptoms returned to or below their pre-FCE level [29].

The observers rated each video twice, in September 2010 (session 1) and in July 2011 (session 2). Between these sessions each rater performed approximately 30 short FCEs (material handling tests only), as part of the regular clinical procedure of a work rehabilitation program. All raters attended both sessions. Data extraction into the database was performed by an individual who was not involved in the data analysis.

Both patients and raters agreed that their data would be used either for the scope of research or education. Because this study was part a regular educational video based training, no ethical approval was required. However, this study was part of a research project approved by the Medical Ethics Committee of Canton Aargau, Switzerland (EK AG 2010/055) [30].

Data analysis

Intra-rater reliability was assessed by comparing the scores from the first rating session with the scores from the second session for each rater. Inter-rater reliability was assessed twice: by comparing the scores between all the raters in session 1 and 2. Category 5 "not classifiable" was excluded from the analyses. Inter-rater and intra-rater reliability was calculated using Cohen's Kappa values for dichotomous data, and squared weighted Kappa values for categorical data and percentages of agreement. A percentage of agreement of 80% or more was judged as acceptable. If agreement was ≥80% and Kappa was κ>0.60 then reliability values were considered as acceptable. [31] AGREE (Agree, Version 7.002) was used to analyze Kappa for multiple observer categories [32] and the ONLINE KAPPA CALCULATOR was used for multiple raters [33]. All other analyses were performed using SPSS (Statistical Package for Social Sciences, Version 20, 2011).

RESULTS

Intra-rater reliability of physical effort determination for all FCE tests

Excluding category 5 "not classifiable" resulted in 325 ratings for the categorical scale for physical effort determination (P_{ED}) (Table 3.1) and 376 ratings were performed for the dichotomous scale for submaximal effort (S_{ED}) (Table 3.2).

Reliability of physical effort determination (P_{ED})

The intra-rater reliability of P_{ED} for all FCE tests in both sessions together was κ=0.49 (95% CI 0.22 to 0.75). The inter-rater agreement of P_{ED} for all FCE tests increased from 73% (session

Table 3.1 Cross tabulation of the categorical ratings for physical effort determination (P_{ED}) in session 1 and 2

Category [a]			Session 2					Total
			1	2	3	4	5	
		Description						
Session 1	1	Light to medium effort	156	32	2	1	4	195
	2	Heavy effort	40	70	5	1	5	121
	3	Maximum effort	2	8	5	0	8	23
	4	Over safe maximum	0	3	0	0	0	3
	5	Not classifiable [b]	7	2	0	0	27	36
Total			205	115	12	2	44	378

[a] Categories 1–5 are described in the Appendices 3.1 and 3.2; [b] Category 5 "not classifiable" was excluded from the analyses.

Table 3.2 Cross tabulation of the categorical ratings for submaximal effort determination scale (S_{ED}) in session 1 and 2

Category [b]		Session 2		Total
		Criteria for maximal physical effort observed [a]		
		Yes	No	
Session 1	Yes	241	27	268
	No	23	85	108
Total		264	112	376

[a] Yes = observed effort was assumed to be indicative for maximal effort as described in Appendices 3.1 and 3.2 when patient performed the material or non-material handling test. [b] No = Submaximal effort was assumed when a patient stopped a material or non-material handling test *before* the FCE rater observed sufficient criteria indicative of maximal weight, or significant functional problems/limitation as described in Appendices 3.1 and 3.2.

1) to 85% (session 2). Kappa values as a measure of inter-rater reliability of P_{ED} for all FCE tests increased from session 1 (0.51; 95% CI 0.23 to 0.80) to session 2 (0.72; 95% CI 0.49 to 0.94). Mean Kappa values for inter-rater reliability of P_{ED} increased from session 1 to 2 for material handling (0.17), postural tolerance (0.21) and ambulation (0.03) (Table 3.3). Mean agreement values of material handling, postural tolerance and ambulation tests ranged from 54% to 75% for inter- and intra-rater reliability (Table 3.3).

Reliability of submaximal effort determination (S_{ED})

For S_{ED} the intra-rater reliability for all FCE tests in both sessions together was κ=0.68 (95% CI 0.60 to 0.76). Kappa values as a measure of inter-rater reliability of S_{ED} for all FCE tests increased from session 1 (0.68; 95% CI 0.60 to 0.76) to session 2 (0.77; 95% CI 0.70 to 0.84). Mean Kappa values for inter-rater reliability of S_{ED} increased from session 1 to 2 for material handling (0.04), postural tolerance (0.47) and ambulation (0.07) (Table 3.3). Mean agreement values of material handling, postural tolerance and ambulation tests ranged from 70% to 97% for inter- and intra-rater reliability (Table 3.3).

Comparison reliability of P_{ED} and S_{ED}

In 6 out of 10 tests inter-rater agreement and Kappa values for the P_{ED} were equal or increased from session 1 to session 2. For S_{ED} inter-rater agreement and Kappa values were similar or increased for all 10 tests. The general reliability of S_{ED} was higher than that of P_{ED}. The inter-rater reliability (% agreement) of S_{ED} was higher in 8 tests (out of 10) for session 1, and in 8 tests (out of 10) for session 2 than that of P_{ED}. The inter-rater reliability (Kappa) of S_{ED} was higher in 7 tests (out of 10) for session 1, and in 8 tests (out of 10) for session 2 than that of P_{ED}. For intra-rater reliability (% agreement/Kappa) S_{ED} was higher than P_{ED} in 10 out of 10 and 5 out of 10 tests respectively. When applying cut-off scores for acceptable reliability (agreement levels ≥80%, κ>0.60), 46% (55 out of 120) of the reliability values fulfilled this criterion (see *italicised* values in Table 3.3).

DISCUSSION

When applying cut-off scores of agreement ≥80%, κ>0.60, the overall reliability of P_{ED} and S_{ED} was acceptable for less than half (46%) of all FCE observations. For S_{ED} reliability was acceptable in the majority (67%) of the FCE tests. However, the reliability of the P_{ED} was acceptable in only 38% of tests. Inter- and intra-rater reliability between each FCE test varied considerably. The increase in mean reliability scores from session 1 to session 2 was on average higher in the P_{ED} than in the S_{ED}.

Table 3.3 Inter- and intra-rater reliability for each FCE test

Category	Test (n)	Physical effort determination scale (P_{ED}) [a]						Submaximal effort scale (S_{ED}) [b]					
			inter		inter	intra			inter		inter	intra	
		Session 1	Session 1	Session 2	Session 2	Session 1-2		Session 1	Session 1	Session 2	Session 2	Session 1-2	
		%	κ	%	κ	%	κ	%	κ	%	κ	%	κ
M	one-handed carrying (4)	68	0.57	*80*	*0.74*	71	0.54	75	0.49	75	0.49	76	0.29
M	lifting floor to waist (4)	58	0.43	73	*0.64*	67	0.47	*85*	*0.70*	*88*	*0.76*	*85*	0.05
M	two-handed horizontal lift (2)	50	0.34	47	0.29	66	0.34	*95*	*0.90*	*100*	*1.00*	*100*	*1.00*
M	lifting waist to overhead (2)	66	0.55	*91*	*0.88*	*81*	0.60	*100*	*1.00*	*100*	*1.00*	*100*	*1.00*
Mean		61	0.47	73	0.64	71	0.49	89	0.77	91	0.81	90	0.59
P	kneeling (1)	*80*	*0.73*	*90*	*0.99*	*84*	-0.08	68	0.35	*100*	*1.00*	*100*	*1.00*
P	forward bend sitting (1)	44	0.25	33	0.11	55	NA	68	0.35	*90*	*0.81*	76	-0.08
P	overhead working (1)	42	0.22	79	*0.72*	50	0.35	74	0.49	*90*	*0.81*	*80*	-0.08
Mean		55	0.40	67	0.61	63	0.14	70	0.40	93	0.87	85	0.28
A	stair climbing (1) [c]	62	0.49	*100*	*1.00*	76	0.00	*90*	*0.80*	*100*	*1.00*	*100*	*1.00*
A	stair climbing (1) [d]	27	0.02	0	-0.33	74	NA	*100*	*1.00*	*100*	*1.00*	*100*	*1.00*
A	walking (1)	73	*0.64*	68	0.57	75	0.14	56	0.12	57	0.14	*90*	*0.76*
Mean		54	0.38	56	0.41	75	0.07	82	0.64	86	0.71	97	0.92

Inter: inter-rater reliability; intra: intra-rater reliability; %: percentage agreement; κ: Cohen's Kappa values for dichotomous, Squared weighted Kappa for categorical data; [a] observational criteria for determination of physical effort during material and non-material handling tests (see Appendices 3.1 and 3.2); [b] submaximal effort was assumed, when a participant stopped a material or non-material handling tests before the FCE rater observed sufficient observational criteria indicative of maximal effort; M: Material handling tests; P: Postural tolerance tests; A: Ambulation tests. (n): number of videos; [c] short video length until patient stops; [d] full video length of the test 10x10 stairs up and down; NA: not applicable, due to lack of cell filling. Italicised values: criteria for acceptable reliability (agreement ≥80%, κ>0.60).

S_{ED} during FCE tests can be reliably detected in the majority of cases. However the results of this study are disappointing, as raters reached the required reliability cut-off values for both the P_{ED} and S_{ED} in less than half of the observations. This finding has clinical relevance for four reasons. First: some FCEs claim to support fitness-for-work determination with an extrapolation of FCE results to job demands [14,34]. The job demands and their frequencies during a working day (occasional, 1–33%; frequent, 34–66%; constant 67–100%) are matched to P_{ED} "maximum", "heavy" and "light to moderate". Good reliability of P_{ED} is needed to enable adequate matching between FCE performance and work demands. Second: FCEs have been reported to accurately describe physical capacity only if a person exerts "maximal" voluntary effort [23,35]. Good reliability of determination of effort is a prerequisite for such a clinical interpretation. Third: FCE reports are used by third parties to inform on the progress of insurance claims. Some interpret submaximal physical effort as 'unmotivated'. The debate over whether this interpretation is valid is beyond the scope of this paper, but it highlights the relevance of the psychometric properties of this determination. Fourth: whether the FCE score represents maximal or submaximal capacity, and the reasons for performing submaximally, are relevant for designing individualized vocational rehabilitation aimed at improvement of functional capacity.

Compared to three previous reliability studies on material handling tests, our values are clearly lower [22,23,26]. In some of these previous studies with high reliability values two-point scales for determination of physical effort were used, which increases the a-priori probability for agreement compared to a multiple item scale as in our study. In our study agreement on the dichotomous scale (submaximal effort determination) was substantially higher too. Moreover the results show on average an increase in the agreement and reliability rating on both the P_{ED} and S_{ED} scales when administered 10 months apart, indicating a "learning" effect. Our data support the assumption that postural tolerance tests may be difficult to rate using the FCE observational methods, but that experience can substantially improve reliability. The average agreement and Kappa values for the inter-rater reliability of P_{ED} increased by 0.40 during the 10-month period. This may be partly attributed to experience. The raters participating in this study used 1-day FCEs for the standard assessment of most in-patients. In addition they received one-to-one supervision from an FCE expert once a year, and their superiors supervised each FCE report as part of regular quality control. Based on the observation in this study that experience and basic training increased reliability scores, we suggest that novice raters using the observational criteria are supervised more intensively than in our study. To what extent observational criteria for effort determination can be improved by additional training remains unknown.

The only slight increase in the agreement and reliability of S_{ED} might be due to the high scores obtained in the first observation session. When tests were grouped according to type of task

the reliability of the physical effort determination scale was generally lower when applied to postural tolerance tests, such as overhead working and kneeling, than when used with material handling tests. This is consistent with results from studies reporting on forward bend, standing and crouching [25,35,36]. Moreover observational criteria seem to be less reliable when applied to ambulation tests such as walking and stair climbing compared to material handling tests [25,36]. However, the results may be influenced by the fact that postural tolerance tests were not part of the regular 1-day FCE utilized in most in-patients, but were only used when indicated. Thus, raters collected more test-experience with the observation of material handling tests than with postural tolerance. Other possible reasons for the lower reliability of the postural tolerance and ambulation tests could be the ceiling effect due to the predefined maximal time limit of the test or the muscular use at submaximal rates. It is theoretically infeasible to judge maximum effort level when submaximal muscular effort is requested e.g. in the overhead work test, the duration of 5 min is not the requested maximum performance, but a time limit. The results of this study underscore this problem. We suggest that observational criteria of physical effort in postural tolerance and ambulation tests need further refinement. To our knowledge no study has been conducted to determine the validity of observational criteria for postural tolerance and ambulation tests in FCE.

In two videos in which a patient performed the one-handed carrying test, ratings showed low agreement. After rating, we discussed these two videos with the raters and asked them where the difficulty lay. Almost half of the raters responded that these were debatable videos due to the pain behavior of the patient. The maximum performance of a patient is determined by the individuals' ability, motivation, and other psychosocial factors [37,38]. However, physical effort determination cannot be used interchangeably with non-organic signs described by Waddell et al., despite some important overlap of the two measurement methods [38]. It has been questioned whether lay persons and health care providers can accurately classify effort during a lifting task performed by actors [39]. Similarly to our results this underscores the challenge of determining effort using a categorical rating scale.

Strengths and weaknesses of the study

The strengths of the study were that the inter- and intra-rater reliability measures were based on the results of a large sample of raters, and multiple observations on patient videos. Compared to most other studies on the reliability of P_{ED}, additionally to the material handling tests, we included postural tolerance and ambulatory tests. Furthermore this is to our knowledge the first study on the reliability of observational criteria used in FCE tests based on two ratings taken within a period of 10 months, excluding the risk of recall bias. We used 18 videos instead of real patients to test the reliability of the observers. The results

may therefore only partly reflect a FCE performed live with the patient. One may argue that several clinical parameters may not have been visible on video tape, such as respiration, and that the raters did not benefit from three-dimensional vision. Observing videos without sound and communication is relevantly different from a clinical setting. In clinical practice FCE raters observe the same patient at different levels of effort when performing the same FCE test. This might facilitate comparison of their own ratings with their previous observations. Studies should be performed to analyze whether the availability of additional information would have changed the results. This study was performed with a sample of four patients. We might therefore not have seen all types of movement patterns of patients with back pain. Because the study was designed to measure the reliability of the raters observing the performance rather than the reliability of that performance, this may have been adequate. The Kappa statistic has an advantage over percentage of agreement because it corrects for chance [31]. In some tests high agreement between raters was observed and Kappa values were in some cases extremely low. This phenomenon may occur when the variation in row and column totals is low [40]. Furthermore it may be debatable if the cut-off score for Kappa values of $\kappa>0.60$ for acceptable reliability used in our study is enough rigorous when one has to make decisions at the individual patient level [41]. The results should therefore be interpreted accordingly. Category 5, "not classifiable", was excluded from the analysis for two reasons. First "not classifiable" relates to another dimension than those categories related to effort. Therefore it cannot be analyzed in the effort domain. Secondly, only a few ratings were "not classifiable", indicating its minor influence.

Future studies

Although there have been some advances in the study of reliability of physical effort determination, major gaps remain: for example, what are valid and practical reference standards for determining maximal physical effort during FCE tests? While some experimental studies measuring muscle activity measurements such as surface EMG, superimposed electrical stimulation, and lactate concentration have been performed, they lack practicality for clinical use [42,43]. How should evidence-based cut-off scores of reliability be defined that are useful for the various purposes of FCE? Future studies should address these unresolved questions and promote the development of a reliable tool for the determination of physical effort, above all for postural tolerance tests.

CONCLUSIONS

The reliability of observing physical effort varied substantially between FCE tests, ranging from unacceptable to good. The dichotomous rating of sub-maximal effort was more reliable than the categorical rating for physical effort determination. However, with both rating scales acceptable reliability values were reached on average only in every second observation, which limits their utility for clinical decision-making. Regular education and training may improve the reliability of observational criteria for effort determination. Further research is needed to develop reliable observation scales.

ACKNOWLEDGEMENTS

The authors thank the physiotherapists of the Department of Work Rehabilitation, Rehaklinik Bellikon who participated in this study. We also thank Doug Gross and Dee Delay for the fruitful discussions on the criteria for physical effort determination. Part of the study was funded by the Swiss Accident Insurance Fund, SUVA (Schweizerische Unfallversicherungsanstalt).

Appendix 3.1 Observational criteria for determination of physical effort during material handling tests

Criteria	Light to moderate	Heavy	Maximum
Muscle recruitment			
Prime movers	Normal recruitment	Bulging	Bulging
Accessory muscles	No or only slight muscle recruitment	Distinct recruitment	Bulging
Base of support	Natural stance	Distinctly increased	Very wide base
Posture	No or only slight counterbalance in extension	Distinctly increased counterbalance	Substantial counterbalance
Heart rate and respiration	No or minimal increases in heart rate and respiration	Distinct increases in heart rate and respiration	Substantial increases in heart rate and respiration
Control and safety	Smooth movements	Increasingly controlled movement; might begin to use momentum; execution with difficulty but not yet at the limit	Still safe but unable to maintain control with the addition of any more weight
Pace	Moderate/ comfortable pace	Distinctly slower; very deliberate movements	Very slow (an increased pace would affect stability and control)

The level of physical effort during material handling tests was determined on the basis of observational criteria indicative of light to moderate, heavy, or maximal weight load [9,18,44]. Maximal effort was assumed when, on the basis of the expertise of the functional capacity evaluation (FCE) rater, sufficient criteria indicative of safe maximal weight were observed. Submaximal effort was assumed when a participant stopped a material handling test before the FCE assessor observed sufficient criteria indicative of maximal weight. Appendix 3.1 is used with permission from Verein IG Ergonomie, Swiss Association of Rehabilitation.

Appendix 3.2 Observational criteria for determination of physical effort during non-material handling tests

Criteria	No or slight functional problem/limitation	Some functional problem/limitation	Substantial functional problem/-limitation
Posture	Maintains normal posture, or slight deviation in posture [a]	Some deviation from normal posture [a], occasional change of position	Substantial deviation from normal posture [a], substantial unrest (frequent change of posture position)
Movement pattern	Normal movement pattern, slight deviation from normal [a], smooth movements or slight muscle stiffness, normal to slightly slower performance	Some deviation from the normal movement pattern [a], tense movements, markedly slower performance	Substantial deviation from the normal movement pattern [a], very tense movements, very slow performance
Muscle recruitment	Normal recruitment of prime movers only, or minimal recruitment of accessory and stabilizing muscles of the trunk, neck or joints stabilizers	Some recruitment of accessory and stabilizing muscles of the trunk, neck or joints stabilizers	Pronounced recruitment of accessory and stabilizing muscles of the trunk, neck or joints
Reaction of the autonomic nervous system	Minimal increase in heart rate	Moderate increase in heart rate and respiration	Substantial increase in heart rate, respiration rate and significant sweating

The level of physical effort during non-material handling tests was determined on the basis of observational criteria indicative of no or slight limitation/problem, some functional limitation/problem, or significant limitation/problem [28]. Maximal effort was assumed when, on the basis of the expertise of the functional capacity evaluation (FCE) rater, sufficient criteria indicative of substantial functional problem/limitation were observed. Submaximal effort was assumed when a participant stopped a non-material handling test before the FCE rater observed sufficient criteria of substantial functional problem/limitation. Appendix 3.2 is used with permission from Verein IG Ergonomie, Swiss Association of Rehabilitation.

[a] Asymmetry (unequal loading) or deviation from neutral.

REFERENCES

1. Gerdle B, Bjork J, Henriksson C, Bengtsson A. Prevalence of current and chronic pain and their influences upon work and healthcare-seeking: a population study. J Rheumatol. 2004;31:1399-406.
2. Bevan S, Quadrello T, McGee R, Mahdon M, Vavrovsky A, Barham L. Fit for Work? Musculoskeletal Disorders in the European Workforce. UK: The Work Foundation, 2009.
3. Lambeek LC, van Mechelen W, Knol DL, Loisel P, Anema JR. Randomised controlled trial of integrated care to reduce disability from chronic low back pain in working and private life. BMJ. 2010;340:c1035.
4. Alschuler KN, Theisen-Goodvich ME, Haig AJ, Geisser ME. A comparison of the relationship between depression, perceived disability, and physical performance in persons with chronic pain. Eur J Pain. 2008;12:757-64.
5. Schiphorst Preuper HR, Reneman MF, Boonstra AM, Dijkstra PU, Versteegen GJ, Geertzen JH. The relationship between psychosocial distress and disability assessed by the Symptom Checklist-90-Revised and Roland Morris Disability Questionnaire in patients with chronic low back pain. Spine J. 2007;7:525-30.
6. Smeets RJ, van Geel AC, Kester AD, Knottnerus JA. Physical capacity tasks in chronic low back pain: what is the contributing role of cardiovascular capacity, pain and psychological factors? Disabil Rehabil. 2007;29:577-86.
7. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain. 2005;113:9-19.
8. Wittink H. Functional capacity testing in patients with chronic pain. Clin J Pain. 2005;21:197-9.
9. Isernhagen SJ. Functional capacity evaluation: rational, procedure, utility of the kinesiophysical approach. J Occup Rehabil. 1992;2:157-68.
10. Henchoz Y, de Goumoens P, Norberg M, Paillex R, So AK. Role of physical exercise in low back pain rehabilitation: a randomized controlled trial of a three-month exercise program in patients who have completed multidisciplinary rehabilitation. Spine (Phila Pa 1976). 2010;35:1192-9.
11. Isernhagen SJ, Galper JS. General testing principles for functional capacity evaluations. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations. Chicago IL: American Medical Association, 2009:41-52.
12. Kool JP, Oesch PR, Bachmann S, Knuesel O, Dierkes JG, Russo M, et al. Increasing days at work using function-centered rehabilitation in nonacute nonspecific low back pain: a randomized controlled trial. Arch Phys Med Rehabil. 2005;86:857-64.
13. Durand MJ, Brassard B, Hong QN, Lemaire J, Loisel P. Responsiveness of the physical work performance evaluation, a functional capacity evaluation, in patients with low back pain. J Occup Rehabil. 2008;18:58-67.
14. Oesch PR, Kool JP, Bachmann S, Devereux J. The influence of a Functional Capacity Evaluation on fitness for work certificates in patients with non-specific chronic low back pain. Work. 2006;26:259-71.

15. Wind H, Gouttebarge V, Kuijer PP, Sluiter JK, Frings-Dresen MH. Complementary value of functional capacity evaluation for physicians in assessing the physical work ability of workers with musculoskeletal disorders. Int Arch Occup Environ Health. 2009;82:435-43.

16. Kuijer PP, Gouttebarge V, Brouwer S, Reneman MF, Frings-Dresen MH. Are performance-based measures predictive of work participation in patients with musculoskeletal disorders? A systematic review. Int Arch Occup Environ Health. 2011;85:109-23.

17. Genovese E, Isernhagen SJ. Approach to requesting a functional evaluation. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations. Chicago IL: American Medical Association, 2009:19-40.

18. Denier-Bont F, Fischer V, Oesch P, Oliveri M. [Functional Capacity Evaluation: Course manual] ed. Bellikon: Verein IG Ergonomie, Swiss Association of Rehabilitation, 2007.

19. Innes E. Handgrip strength testing: A review of the literature. Aust Occup Ther J. 1999;46:120-40.

20. Gouttebarge V, Wind H, Kuijer PP, Frings-Dresen MH. Reliability and validity of Functional Capacity Evaluation methods: a systematic review with reference to Blankenship system, Ergos work simulator, Ergo-Kit and Isernhagen work system. Int Arch Occup Environ Health. 2004;77:527-37.

21. Reneman MF, Fokkens AS, Dijkstra PU, Geertzen JH, Groothoff JW. Testing lifting capacity: validity of determining effort level by means of observation. Spine (Phila Pa 1976). 2005;30:E40-6.

22. Smith RL. Therapists' ability to identify safe maximum lifting in low back pain patients during functional capacity evaluation. J Orthop Sports Phys Ther. 1994;19:277-81.

23. Gardener L. Reliability of occupational therapists in determining safe, maximal lifting capacity. Aus Occup Ther J. 1999;46:110-9.

24. Jay MA, Lamb JM, Watson RL, Young IA, Fearon FJ, Alday JM, et al. Sensitivity and specificity of the indicators of sincere effort of the EPIC lift capacity test on a previously injured population. Spine (Phila Pa 1976). 2000;25:1405-12.

25. Brouwer S, Reneman MF, Dijkstra PU, Groothoff JW, Schellekens JM, Goeken LN. Test-retest reliability of the Isernhagen Work Systems Functional Capacity Evaluation in patients with chronic low back pain. J Occup Rehabil. 2003;13:207-18.

26. Isernhagen SJ, Hart DL, Matheson LM. Reliability of independent observer judgments of level of lift effort in a kinesiophysical Functional Capacity Evaluation. Work. 1999;12:145-50.

27. Gross DP, Battie MC. Reliability of safe maximum lifting determinations of a functional capacity evaluation. Phys Ther. 2002;82:364-71.

28. Work Well Systems Inc. Functional Capacity Evaluation V.2.ed. Duluth MN: Work Well Systems Inc, 2006.

29. Reneman MF, Kuijer W, Brouwer S, Preuper HR, Groothoff JW, Geertzen JH, et al. Symptom increase following a functional capacity evaluation in patients with chronic low back pain: an explorative study of safety. J Occup Rehabil. 2006;16:197-205.

30. Trippolini MA, Reneman MF, Jansen B, Dijkstra PU, Geertzen JH. Reliability and safety of functional capacity evaluation in patients with whiplash associated disorders. J Occup Rehabil. 2013;23:381-90.

31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74.

32. Popping R. AGREE, a package for computing nominal scale agreement. Comput Stat Data Anal. 1984;2:182-5.

33. Randolph JJ. Online Kappa Calculator, 2008. Available from: http://justus.randolph.name/kappa. Accessed June 21, 2012.

34. Wind H, Gouttebarge V, Kuijer PP, Sluiter JK, Frings-Dresen MH. Effect of Functional Capacity Evaluation information on the judgment of physicians about physical work ability in the context of disability claims. Int Arch Occup Environ Health. 2009;82:1087-96.

35. Gouttebarge V, Wind H, Kuijer PP, Sluiter JK, Frings-Dresen MH. Intra- and interrater reliability of the Ergo-Kit functional capacity evaluation method in adults without musculoskeletal complaints. Arch Phys Med Rehabil. 2005;86:2354-60.

36. Durand MJ, Loisel P, Poitras S, Mercier R, Stock SR, Lemaire J. The interrater reliability of a functional capacity evaluation: the physical work performance evaluation. J Occup Rehabil. 2004;14:119-29.

37. van Abbema R, Lakke SE, Reneman MF, van der Schans CP, van Haastert CJ, Geertzen JH, et al. Factors associated with functional capacity test results in patients with non-specific chronic low back pain: a systematic review. J Occup Rehabil. 2011;21:455-73.

38. Oesch P, Meyer K, Jansen B, Mowinckel P, Bachmann S, Hagen KB. What is the role of "nonorganic somatic components" in functional capacity evaluations in patients with chronic nonspecific low back pain undergoing fitness for work evaluation? Spine (Phila Pa 1976). 2012;37:E243-50.

39. Schapmire DW, St James JD, Townsend R, Feeler L. Accuracy of visual estimation in classifying effort during a lifting task. Work. 2011;40:445-57.

40. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43:543-9.

41. Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. 4th ed. Oxford: Oxford University Press, 2008.

42. Moesch W. The use of objective parameters for functional capacity evaluations. FCE. Faculty of Medicine. Göttingen: Georg-August-Universität zu Göttingen, 2005:70.

43. Verbunt JA, Seelen HA, Vlaeyen JW, Bousema EJ, Van Der Heijden GJ, Heuts PH, et al. Pain-related factors contributing to muscle inhibition in patients with chronic low back pain: An experimental investigation based on superimposed electrical stimulation. Clinical Journal of Pain. 2005;21:232-40.

44. Oesch P, Meyer K, Bachmann S, Hagen KB, Vollestad NK. Comparison of two methods for interpreting lifting performance during functional capacity evaluation. Phys Ther. 2012;92:1130-40.

Chapter 4

Reliability and safety of functional capacity evaluation in patients with whiplash-associated disorders

Maurizio A. Trippolini
Michiel F. Reneman
Beatrice Jansen
Pieter U. Dijkstra
Jan H. B. Geertzen

Journal of Occupational Rehabilitation. 2013; 23: 381-390.

ABSTRACT

Introduction: Whiplash-associated disorders (WAD) are a burden for both individuals and society. It is recommended to evaluate patients with WAD at risk of chronification to enhance rehabilitation and promote an early return to work. In patients with low back pain (LBP), Functional Capacity Evaluation (FCE) contributes to clinical decisions regarding fitness-for-work. FCE should have demonstrated sufficient clinimetric properties. Reliability and safety of FCE for patients with WAD is unknown.

Methods: Thirty-two participants (11 females and 21 males; mean age 39.6 years) with WAD (Grade I or II) were included. The FCE consisted of 12 tests, including material handling, hand grip strength, repetitive arm movements, static arm activities, walking speed, and a 3 min step test. Overall the FCE duration was 60 minutes. The test-retest interval was seven days. Interclass correlations (model 1) (ICCs) and limits of agreement (LoA) were calculated. Safety was assessed by a Pain Response Questionnaire, observation criteria and heart rate monitoring.

Results: ICCs ranged between 0.57 (3 min step test) and 0.96 (short two-handed carry). LoA relative to mean performance ranged between 15% (50 m walking test) and 57% (lifting waist to overhead). Pain reactions after WAD FCE decreased within days. Observations and heart rate measurements fell within the safety criteria.

Conclusions: The reliability of the WAD FCE was moderate in two tests, good in five tests and excellent in five tests. Safety-criteria were fulfilled. Interpretation at the patient level should be performed with care because LoA were substantial.

INTRODUCTION

Whiplash injuries occur primarily after motor vehicle collisions, but can also occur during work, sports or other mishaps leading to an indirect cervical trauma. The Québec Task Force (QTF) on Whiplash-Associated Disorders (WAD) defined whiplash as "an acceleration-deceleration mechanism of energy transferred to the neck that results in soft tissue injury that may lead to a variety of clinical manifestations including neck pain and its associated symptoms" [1]. Patients with WAD may also suffer from upper limb pain, paresthesias, psychological distress, anxiety, dizziness, headache, fatigue, nausea, concentration deficits and many more symptoms [2,3]. WAD refers to the clinical entities related to the injury, but should be distinguished from the injury mechanism [1].

Whiplash injury incurs large economic, social and personal burden. Recent studies report that 10-40% of patients with WAD will fail to recover [1,4,5]. If recovery occurs, this will take place within the first 2-3 months [6]. The WAD Task Force proposed that patients with WAD who do not return to work within 6 to 12 weeks after injury receive an interdisciplinary assessment including disability measures so that interventions may be specifically directed, potentially averting the course to chronicity [7,8].

Functional Capacity Evaluations (FCE) were developed to assess work-related abilities [9,10]. These work-related tests were based on a taxonomy described in the US Department of Labor's Dictionary of Occupational Titles (DOT) [11]. Although no consensus on the definition of FCE is available [12], we use the term as follows: FCE is an evaluation of the capacity to perform activities that is used to make recommendations for participation in work while considering the individuals' body functions and structures, environmental factors, personal factors and health status. During the past decade, measurement properties of FCEs such as reliability, validity and safety have been determined [13]. However, these measurement properties have mainly been investigated in patients with low back pain [14] and, to a lesser extent, in healthy subjects [15] and patients with the early stages of osteoarthritis of knees and hips [16], work-related upper limb disorders [17], and work-related neck disorders [18]. Moreover, there is a lack of knowledge on measurement error of FCE, which seriously limits clinical decision making. Furthermore it has been proposed to perform FCE in a more specific and efficient way by selecting a limited number of activities targeted to the workers condition [19,20]. To date no specific FCE for WAD has been developed. The safety of work-related assessments has been recognized as a necessary attribute of FCE studies [21,22], but safety issues such as pain-reaction, muscle soreness, adverse effects and pain medication use have not been reported in patients with WAD.

Hence the aim of this study was to analyze test-retest reliability, measurement error and safety of FCE in patients with WAD who did not return to work within 6 to 12 weeks after injury and who received workers' compensation.

METHODS

Participants

Participants from all over the country (Switzerland) were referred by either a physician or a case manager of the worker's compensation insurance for a half-day comprehensive interdisciplinary rehabilitation assessment. Participants were referred when they had not regained full working capacity within 6 to 12 weeks after initial injury. From January to October 2011, n=71 patients, with WAD were asked to participate in this reliability study after they had completed their FCE. Inclusion criteria were if participants had symptoms according the Québec Task Force-Classification of WAD, grade I (pain, stiffness or tenderness without physical signs) or grade II (pain, stiffness, or tenderness with decreased range of motion and point tenderness), main pain in the head or neck region, sufficient German language skills to communicate with the assessors (all questionnaires were available in German and five foreign languages spoken by the participants), an age of 18–65 years, and willingness to participate (signed informed consent). Exclusion criterion was co-morbidity which considerably limited function, such as neurological deficits, rheumatoid diseases, spinal fractures, tumors, osteoporosis, psychiatric disorders, pregnancy, cardiac hypertension etc. Based on convenience, a sample of participant was selected by an independent person, not involved in the testing procedure, to participate in the retest. In the recruitment period 75 patients fulfilled the inclusion criteria and were asked by an independent person, not involved in the testing procedure, to participate in the reliability study.

A convenience sample of 4 physiotherapists (2 female, 2 male) conducted the FCEs. All attended the official two day FCE training course, are accredited as FCE-Therapists by the Swiss Association of Rehabilitation [23], had performed at least 20 WAD FCEs in the previous year (median 28, interquartile range (IQR), 21 to 37) and had a minimum of 2 years working experience in vocational rehabilitation (median 7 years, IQR 2 to 14), and a minimum professional practice experience of 2 years (mean 14 years, IQR 4 to 23). For this study, all physiotherapists received an additional half day training, and had a WAD FCE supervised by an FCE expert.

Procedure

All participants received written and verbal information about the study. Participants were informed that they would be allowed to withdraw their participation at any time without disclosing reasons and without consequences for their medical care. The Medical Ethics

Committee of Canton Aargau granted the ethical approval for this study (EK AG 2010/055). Participants received reimbursement of travel expenses and 50 Swiss francs after completion of the second FCE session.

Study design

A test-retest design was used. During the first visit a review of the medical history and a physical examination was performed by a physician lasting approximately 60 minutes, followed by WAD FCE administered by a physiotherapist. Administration of the WAD FCE lasted approximately 60 minutes.

After the first FCE participants were asked whether he would want to participate in a retest. The fixed order of the tests was standardized and constant between sessions. The second WAD FCE was conducted one week later (median 7 days, IQR 6 to 7). This time period between the two tests needed be long enough to reduce carry-over effects and delayed muscle soreness [24], and short enough considering that the health condition of the study participants may still change. The second WAD FCE was administered by the same tester. Time and day for the retest session were held constant as much as possible. Participants and testers were blinded to the results of the first WAD FCE.

Measures

Functional Capacity Evaluation

The FCE applied in this study (WAD FCE) consisted of 12 tests, based in part on the WorkWell FCE (formerly the Isernhagen Work System) [25]: handgrip strength (left and right), lifting floor to waist, lifting waist to overhead, short two-handed carry, long right- and left-handed carry, overhead work, repetitive reaching (left to right and right to left [17], 50 m walking test [26] and a 3 min step test [27]. Test descriptions are presented in the Appendix. Participants were briefly instructed on how to perform each test. The evaluator first gave a single demonstration of each test. Participants were then asked to perform the tests to their maximum ability. Weights lifted were gradually increased according to a participant's performance, using weights of 2.5 and 5 kilograms (kg). To determine the physical effort level, testers used observational critera [23,25]. Testing could be terminated for four reasons: the participant stopped because of, for example, pain; the observer deemed testing to have become unsafe based on biomechanical criteria; heart rate exceeded 85% of the age-related maximum (220 minus age of participant); or a predefined time limit was reached.

Safety

Safety of the FCE was assessed by heart rate monitoring, observational criteria for effort level during work related tasks, pain reaction as measured with the Pain Response Questionnaire (PRQ) [24], additional pain medication, or reports of serious adverse effects . Participants were asked to score their pain for 17 separate body regions in an 11-point NRS, in which 0 was "no pain" and 10 was "worst pain". Participants were also asked whether their pain was attributable to muscle soreness, to a different origin, a combination of these, or of unknown origin. The participants were asked to fill in the PRQ on the subsequent days (using a diary) after the first WAD FCE until the day of the retest. The WAD FCE was considered safe under the following conditions: when the heart rate did not exceed the age-related maximum, when it did not exceed the maximum observational criteria for effort level during work-related tasks, when it did not lead to injuries, when it resulted in no serious adverse effects, when it did not increase by more than three NRS points [28], and when reported muscle soreness increased in the first 24–48 hours (which is a normal response), subsided during the following two days and then returned to pretest levels within 5–7 days [24]. A response which did not adhere to this definition was interpreted as an abnormal response.

Additional measures

Participant characteristics included age, gender, marital status, education, nationality, work status, current litigation, and compensation-status, among others. Pain intensity was measured with an 11-point numeric rating scale (NRS) [29].

Disability. Neck pain-related disability was measured with the Neck Disability Index (NDI) [30]. The NDI contains 10 items, ranging from no disability (0) to total disability (5). The maximal overall score is 50 points (complete disability).

Anxiety and depression. Anxiety and depression were measured using the Hospital Anxiety and Depression Scale (HADS) [31]. The HADS consists of two scales, one for anxiety and one for depression. Each scale contains seven items, with each item rated from 0 (best) to 3 (worst). The scale scores are calculated by summing the responses to the items up to a maximum score of 21 points per scale (severe case) [32].

Self-efficacy. Self-efficacy in execution of activities which involve the spine was measured with the Spinal Function Sort (SFS) [33]. The instrument contains 50 drawings of activities that involve the spine with simple descriptions. Participants rated self-efficacy for each activity from "able" (4) to "unable" (0). The SFS yields a single rating ranging from 0 to 200.

Data analysis

Depending on data-distribution, test and retest data were analyzed using parametric or non-parametric statistics. Test-retest reliability was expressed as an Interclass Correlation Coefficient (model 1; one-way random) (ICC). ICC was interpreted as follows: ICC≥0.90 is excellent; good when ICC was between 0.75 and 0.90; moderate when ICC was between 0.50 and 0.75; and poor when ICC≤0.50. ICCs were acceptable when ICC≥0.75, and the lower boundary of the 95% confidence interval of the ICC≥0.50 [34]. Agreement was expressed in limits of agreement (LoA) (mean difference ± 1.96 x standard deviation of mean difference) [35]. The ratio between the limits of agreement and the mean score of two sessions was calculated (LoA/mean of two sessions) × 100%), to determine the relative width of the limits of agreement, and to allow comparison of LoA to other studies. Paired t-tests were used to analyze systematic differences between the first and second test session. A response which did not accord to this definition was interpreted as an abnormal response. An analysis was performed to identify differences between those participants who completed two sessions and those who only completed one session. All analyses were performed with SPSS (Statistical Package for Social Sciences, Version 19).

RESULTS

Of the eligible participants, 32 (45.1%) completed both sessions, and 39 (54.9) did not participate in the retest. The reasons for not participating were as follows: 21 (54%) of participants were working at the time of the retest; 6 (15%) explicitly did not want to participate with no reason declared; 4 (10%), did not feel capable due to temporary pain increase at the time of the first WAD FCE; and 8 (21%) mentioned other reasons, such as being on holiday, no transport facilities available etc. A total of 32 participants performed all of the tests. Demographic and clinical variables of the study sample are presented in Table 4.1. The four physiotherapists conducted between 6 and 11 WAD FCEs each.

Reliability and agreement

ICC ranged between 0.54 and 0.96 (Table 4.2). Ratios of the LoA of the WAD FCE tests were between 15% (50 m walking test) and 57% (lifting waist to overhead). Bland and Altman plots revealed variances that were not related to the magnitude of the outcome (plots not shown). The mean performance of the participants increased in the second session in 8 WAD FCE tests, of which three were statistically significant results ($p<0.05$).

Table 4.1 Participants characteristics (n=32)

	Mean (SD)
Age (years)	39.6 (12.3)
BMI	28.2 (5.4)
Disability (NDI 0-50)	21.7 (5.8)
Anxiety (HADS 0-21)	7.3 (4.3)
Depression (HADS 0-21)	6.1 (3.6)
Self-efficacy (SFS 0-200)	146.4 (31.6)
Injury duration since (days), SD	89.6 (33.9)
	n or %
Work capacity for the own job (in %) at the time of WAD FCE*	62.8% (38.5)
Gender: female	11 (34%)
Marital status: married	9 (28%)
Nationality: Swiss	22 (69%)
Education	
Low**	10
Intermediate	21
High	1
Physical work demands[+]	n
0–10 kg	10
11–25 kg	8
25–50 kg	9
>50 kg	5

* work capacity was assessed by the referring physician, ** low = no vocational education, intermediate = vocational education, high = bachelor or higher education. [+] Physical work demands according to the Dictionary of Occupational Titles (DOT). BMI = Body mass index formula: weight (kg) / height $(cm)^2$; NDI = Neck Disability Questionnaire; HADS = Hospital Anxiety Depression Scale; SFS = Spinal Function Sort.

Safety

Except for one participant who had to stop the material handling test because his/her heart rate reached in excess of 85% maximum, all the WAD FCE tests were completed before the 85% maximum heart rate was reached. At the endpoint of each of the material handling tests of the first test session, the mean heart rate difference to the theoretical age-related maximum was 35.9 (SD 16.6). The mean NRS pain before the first WAD FCE was 4.3 (1.8), and 5.3 (SD 1.9) after WAD FCE, p-value <0.001, (mean change -1.1, SD change 1.3). For the second WAD FCE session, these values were 4.3 (SD 1.9) for NRS pain before and 4.9 (SD 1.8) for NRS pain after, p-value <0.001 (mean change 0.6, SD change 1.1). On an individual level, pain increased by two or more NRS points in 18 participants (57%), with none exceeding

Table 4.2 Test results of 2 WAD FCE sessions, and limits of agreement and intra class correlation between the test results

WAD FCE items	Mean session 1	SD session 1	Mean session 2	SD session 2	Mean difference	SD of mean difference	95% CI of mean difference	p	LoA	Ratio of LoA (%)	ICC	95 % CI of ICC	Interpretation of ICC
Hand grip strength right (kgF)	37.8	14.6	40.5	14.3	-2.6	5.3	-4.5 to -0.7	.008	-13.0 to 7.7	26	0.92	0.84 to 0.96	Excellent
Hand grip strength left (kgF)	35.0	14.7	38.7	13.6	-3.6	5.8	-5.8 to -1.5	.001	-15.1 to 7.8	31	0.89	0.78 to 0.94	Good
Lifting floor to waist (kg)	24.1	9.7	24.7	8.9	-0.6	3.8	-2.0 to 0.7	.354	-8.0 to 6.7	30	0.92	0.84 to 0.96	Excellent
Lifting waist to overhead (kg)	13.8	5.8	15.3	4.9	-1.5	4.2	-3.0 to 0.0	.054	-9.8 to 6.8	57	0.66 *0.80	0.42 to 0.82 *0.64 to 0.90	Moderate *Good
Short carry two handed (kg)	32.5	12.7	33.2	13.2	-0.7	3.6	-2.0 to 0.6	.288	-7.7 to 6.4	21	0.96	0.92 to 0.98	Excellent
Long carry right handed (kg)	20.9	6.9	20.2	6.5	0.6	3.0	-0.5 to 1.7	.255	-5.3 to 6.6	29	0.90	0.80 to 0.95	Good
Long carry left handed (kg)	19.4	6.3	19.5	5.7	-0.1	2.6	-1.1 to 0.8	.759	-5.2 to 4.9	26	0.91	0.83 to 0.96	Excellent
Overhead working (sec)	223.0	97.3	227.7	90.9	-4.7	55.8	-24.8 to 15.4	.636	-114.1 to 104.7	49	0.83	0.68 to 0.91	Good
Repetitive reaching right (sec)	77.2	24.5	72.0	21.2	5.3	13.6	0.3 to 10.2	.037	-21.5 to 32.0	36	0.81	0.64 to 0.90	Good
Repetitive reaching left (sec)	77.1	24.5	72.6	22.0	4.5	14.0	-0.5 to 9.5	.078	-22.9 to 31.9	37	0.81	0.65 to 0.90	Good
50 m walking test (km/h)	5.1	0.9	5.2	1.0	-0.1	0.4	-0.2 to 0.1	.362	-0.9 to 0.7	15	0.91	0.82 to 0.95	Excellent
3 min step test (mean heart rate after 1st minute)	116.8	29.7	116.8	20.7	9.0	24.0	-8.7 to 8.9	.988	-46.9 to 47.0	40	0.57	0.28 to 0.77	Moderate

SD = standard deviation; LoA = limits of agreement; 95% CI = 95% confidence interval; ICC = interclass correlation coefficient. Ratio of LoA (%): the ratio between the limits of agreement and the mean score. ((1.96 x standard deviation of mean difference) / mean session 1 and 2 × 100%).
* Results of an analysis when 1 participant who refused to lift any weight in the first session, was excluded from the analysis (see discussion).

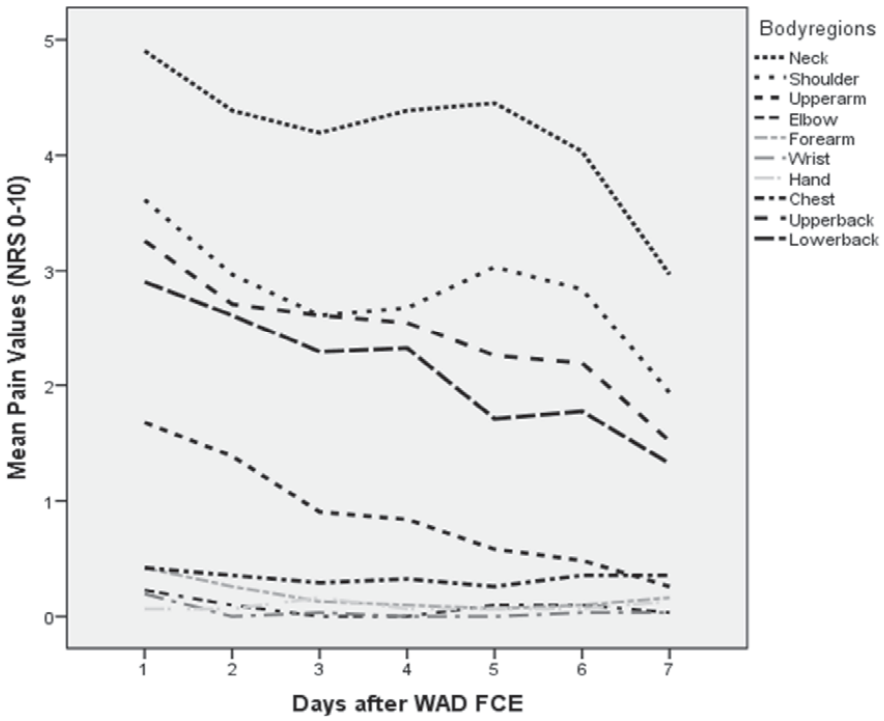


Figure 4.1 Means of the reported pain response per day after WAD FCE measured by the pain response questionnaire (PRQ).

three points. Symptoms also decreased to a mean at pre-test levels in seven days. Average pain scores in the neck and shoulder region measured with the PRQ decreased after the second day post WAD FCE (Figure 4.1). One participant did not complete the PRQ and was excluded. No serious adverse events were reported during or after test and retest.

On average non-participants performed less than participants. We performed a Mann-Whitney U Test for independent-samples to compare the WAD FCE results of the first session between the group that was retested and the group that was not retested (non-participants). In nine out of 12 WAD FCE tests, the results showed no significant difference between the groups. In the three WAD FCE tests with the significantly different test results, the non-participants lifted less in a short two handed carry test (Mean 24.4 kg, SD 12.7), and in the long carry one handed test (Mean right: 16.9 kg SD 7.7; Mean left: 16.3 kg, SD 7.4). Additionally, we compared clinical characteristics, such as neck pain disability, anxiety, depression levels, self-efficacy and pain scores. These characteristics did not differ significantly between participants and the non-participants.

DISCUSSION

Reliability

The test-retest reliability out of 11 to 12 WAD FCE test items was good to excellent. Healthy volunteers [15], patients with chronic low back pain [14] or patients with osteoarthritis of hip and / or knee [16] showed smaller variability in this FCE test compared to the WAD FCE. The following reasons may explain these differing results. In the case of healthy volunteers, who are less affected by pain, less variability in the test results is expected. FCE in the capacity of a patient with chronic low back will not change between two sessions because they are in a relatively stable i.e. chronic phase of the illness. The study of osteoarthritis patients [16] involved conducting the retest study one day after the first test session, therefore a lower variability may be explained by recall bias due to the limited time between the two test sessions. As expected from WAD patients suffering from pain in the neck region, larger LoA scores were observed in the tests affecting the upper body regions i.e. "overhead work" and "lifting waist to overhead".

Lifting from waist to overhead had a moderate ICC (0.66), with significantly different values recorded between the first and second session. This result was in part due to a participant who refused to lift any weight overhead in the first session, but lifted 15 kg in the second session. A post-hoc sensitivity analysis was performed by excluding that participant from the analysis. The ICC value then increased to 0.80, which indicated good reliability.

Regarding the overhead work test with an ICC of 0.83, the larger LoA ratios may also be partly explained by the longer duration of the test at 5 min, compared to the maximum of 90 s in the material handling tests. The longer a test, the greater the chance that the patient would perform differently in another test session. For example, in the study of Brouwer et al. [14], the reliability expressed as an ICC of a 15 min overhead work test was 0.36. To prevent ceiling effects, other researchers have modified the overhead work test by having the patients wear two cuff-weights of 1 kg around their forearm [36]. This procedure results in a reduction of endurance in the overhead work in healthy participants, and an ICC of 0.90 [17]. The results of the hand grip force (in position 2 of the Jamar hand dynamometer) proved to have good to excellent reliability, similarly to the findings of previous studies on hand grip force [37], underlining its clinical use in the evaluation of grip strength in several musculoskeletal disorders. In the repetitive reaching test, ICC values were slightly higher in WAD patients when compared to healthy participants, while LoA were between -21.5 and 32.0 in WAD patients and -9.0 to 12.6 in healthy participants [17]. Tests results of the 3 min step test and 50 m walking test did not change significantly between the two sessions compared to the materials handling tests. It is very unlikely that endurance and gait speed

would improve in that length of time between the two sessions. Our participants were a sample of patients with sub-acute WAD, whose health status was still subject to possible change (improvement). The time interval between the two sessions therefore had to be far enough apart to avoid fatigue, learning or memory effects, but not too far apart to allow a change in health status. We therefore chose a time interval of seven days to take these factors into account. This time period was shorter than previous reliability studies, which had time intervals of 10 to 21 days [14,17,38]. Clinically the measurement error of the test under investigation lies within ± 95% LoA. This means that, at the individual level, a patient's performance could be considered to be changed when it exceeded the LoA. For example in "lifting floor to waist", a patient's performance improved if his performance increased by more than 6.7 kg.

Large limits of agreement scores in health outcome measure are common in pain patients [33,39,40]. As already stated there are no cut-off points of LoA [41]. However one study from Keller et al. who calculated the LoA for the Astrand bicycle test and other back strength tests in LBP patients judged a test with LoA of ≥42% as unreliable [42]. Based on this arbitrary cut-off value, 2 out of the 12 tests of the WAD FCE would be classified as unreliable. This large within-patient variance may be attributed to measurement and random errors of test procedure, evaluator inconsistencies, and patient behavior being influenced by motivation or pain. As hypothesized by others [14,43], but not tested in this study, we argue that a large part of the variance can be attributed to variation within the patients.

Safety

In a Delphi Survey of FCE experts, safety was defined as: "a situation that, given the known characteristics of the person, the procedure should not be expected to lead to injury" [12]. We controlled for safety by using self-report measures such as the NRS, with a diary questionnaire, the PRQ, and measurements taken by the physiotherapist (e.g. heart rate, observation criteria). Based on our results of the PRQ, as reported in Figure 4.1, we conclude that the WAD FCE temporarily increased pain at a similar rate to healthy volunteers [24] and patients with low back pain following FCE [21]. Similarly to both other studies, symptoms in WAD patients also decreased within a week. No safety problems were encountered, and heart rate increased only moderately, with only one patient reaching the 85% heart rate limit in the WAD FCE tests. 10 patients refused to participate, 4 because of pain increase and 6 for unknown reasons. None of these, nor any other participant, reported a formal complaint and no serious adverse effects were reported. We therefore believe that safety was not compromised.

Limitations and strengths of the study

A limitation of this study was that only 45% of the eligible 72 participants were willing to participate in the second session. The main reason was: lack of time (most were already returned to work, others were on holiday, or were living a long distance away etc.). The same phenomenon was found in a FCE test-retest study of Brouwer et al. were approximately 100 patients were eligible during one year, but only 30 patients were willing to participate [14]. In most instances, reasons for not participating were that testing would take too much time, which is similar to the Brouwer et al. study. It is unknown how non-participants would have influenced reliability of the WAD FCE tests. As learning effects influence test-retest reliability [44,45], we did not inform participants of the detailed test results, and ensured the memory effect was minimized by maintaining a large enough time interval between test occasions. Additionally, all test protocols from the first session were collected immediately after the test procedure by an independent person, who was not involved in the testing procedure. Test protocols remained inaccessible for the testers involved. Results of paired t-tests between the two test occasions showed a general trend towards a slightly increased performance on the second occasion. This is in line with test results of healthy volunteers, who scored on average higher on the second test session [15,17]. Although we did not expect test effects such as increased strength and mobility after the first testing session, other effects, such as increased self-efficacy, reassurance etc., may have occurred, creating consistent change within participants. Such a systematic effect will not necessarily affect reliability coefficients [44].

In our study 30% of non-native Swiss patients participated in the study, which is a slight overrepresentation compared to the general Swiss population with 23% with non-native citizens [46]. This is in contrast to previous FCE reliability studies [14,16,38] where mainly native citizens participated. Results of interventions may vary considerably between native and non-native patients [47], but to our knowledge, this has never been the subject of a study in a setting similar to ours (performance testing, reliability, agreement, safety). We therefore think that the results, although taken from a small study sample, might support the utility of the WAD FCE in non-native patients.

Secondly our testers were selected from a sample of 24 physiotherapists. The range of clinical experience covered a wide range of experience (from very low to extensive) encountered in clinical daily practice. Contrary to previous reliability studies where very experienced clinicians performed the FCE tests [6,16,37], our sample of assessors covered a wider range of working experience and age. This might strengthen the generalizations of the results of this study. Our study was conducted in a "real world" environment where patients with delayed recovery were sent to the WAD FCE, compared to some previous FCE reliability studies based on video analysis [43,48].

Participants were referred by physicians and case managers from the German speaking part of Switzerland; to what extent this referral resulted in a population different from other WAD populations is unknown. Because the clinical characteristics of the non-participants did not differ from the participants, nor did the majority of test results, we assume that the selection procedure did not introduce bias relevant for the outcomes of this study (i.e. reliability, agreement, safety). Since the majority of WAD patients are suffering from WAD Grade 1 and 2 [49] , the results of this study may be applied to patients with WAD Grade 1 and 2 who are still suffering from WAD 9–12 weeks after injury and are not working due to WAD.

CONCLUSION

In conclusion, we observed a good to excellent test-reliability in the majority of the WAD FCE tests, while safety-criteria were fulfilled. Clinical interpretation at the individual patient level should be performed with care, however, because of the large LoA.

ACKNOWLEDGEMENTS

The authors thank the physiotherapists Yves Weder, Nicole Saghy-Steger, and the physicians of the Department of Work Rehabilitation, Rehaklinik Bellikon, for their help in recruiting participants for this study. We also thank Peter Erhart, Claudia Diethelm, Axel Gehrke for data preparation, technical and administrative support, and all participants for their participation. We are very grateful to Michael Oliveri, Hans Peter Gmünder, Thomas Mäder, Felix Weber, Salih Muminagic and Sönke Johannes who served as guarantors for the project.

This study was supported by the Rehaklinik Bellikon and the Swiss Accident Insurance Fund (SUVA). No benefits in any form have been or will be received from a commercial party related directly or indirectly to the subject of this manuscript.

REFERENCES

1. Spitzer WO, Skovron ML, Salmi LR, Cassidy JD, Duranceau J, Suissa S, et al. Scientific monograph of the Quebec Task Force on Whiplash-Associated Disorders: redefining “whiplash” and its management. Spine (Phila Pa 1976). 1995;20:1S-73S.
2. Ferrari R, Russell AS, Carroll LJ, Cassidy JD. A re-examination of the whiplash associated disorders (WAD) as a systemic illness. Ann Rheum Dis. 2005;64:1337-42.
3. Radanov BP, di Stefano G, Schnidrig A, Ballinari P. Role of psychosocial stress in recovery from common whiplash [see comment]. Lancet. 1991;338:712-5.

4. Chappuis G, Soltermann B. [Accident rates and Costs of Whiplash Associated Disorders. A Swiss pecularity?]. Rev Med Suisse. 2006;6:398-406.

5. Holm LW, Carroll LJ, Cassidy JD, Hogg-Johnson S, Cote P, Guzman J, et al. The burden and determinants of neck pain in whiplash-associated disorders after traffic collisions: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine (Phila Pa 1976). 2008;33:S52-9.

6. Kamper SJ, Rebbeck TJ, Maher CG, McAuley JH, Sterling M. Course and prognostic factors of whiplash: a systematic review and meta-analysis. Pain. 2008;138:617-29.

7. Strebel H, Ettlin T, Annoni J, Caravatti M, Jan S, Gianella C, et al. [Diagnostics and therapeutic approach in the acute phase of whiplash associated disorders. Guidelines of the Swiss WAD Task Force] Schweiz Med Forum. 2002;2:119-25.

8. Stöckli H, Ettlin T, Gysi F, Knüsel O, Marelli R, B S. [Diagnostics and therapeutic approach in the chronic phase of whiplash associated disorders]. Schweiz Med Forum. 2005;5:1182-7.

9. Genovese E, Galper JS. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations. ed: American Medical Association, 2009.

10. Kuijer PP, Gouttebarge V, Brouwer S, Reneman MF, Frings-Dresen MH. Are performance-based measures predictive of work participation in patients with musculoskeletal disorders? A systematic review. Int Arch Occup Environ Health. 2011;85:109-23.

11. U.S. Department of Labor. The Revised Handbook for Analyzing Jobs. 4th ed. Indianapolis: JIST Works, inc., 1991.

12. Soer R, van der Schans CP, Groothoff JW, Geertzen JH, Reneman MF. Towards consensus in operational definitions in functional capacity evaluation: a Delphi Survey. J Occup Rehabil. 2008;18:389-400.

13. Reneman M, Wittink H, Gross DP. The scientific status of functional capacity evaluation. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations: American Medical Association, 2009:393-420.

14. Brouwer S, Reneman MF, Dijkstra PU, Groothoff JW, Schellekens JM, Goeken LN. Test-retest reliability of the Isernhagen Work Systems Functional Capacity Evaluation in patients with chronic low back pain. J Occup Rehabil. 2003;13:207-18.

15. Reneman MF, Brouwer S, Meinema A, Dijkstra PU, Geertzen JH, Groothoff JW. Test-retest reliability of the Isernhagen Work Systems Functional Capacity Evaluation in healthy adults. J Occup Rehabil. 2004;14:295-305.

16. van Ittersum MW, Bieleman HJ, Reneman MF, Oosterveld FG, Groothoff JW, van der Schans CP. Functional capacity evaluation in subjects with early osteoarthritis of hip and/or knee; is two-day testing needed? J Occup Rehabil. 2009;19:238-44.

17. Soer R, Gerrits EH, Reneman MF. Test-retest reliability of a WRULD functional capacity evaluation in healthy adults. Work. 2006;26:273-80.

18. Reesink DD, Jorritsma W, Reneman MF. Basis for a functional capacity evaluation methodology for patients with work-related neck disorders. J Occup Rehabil. 2007;17:436-49.

19. Gouttebarge V, Wind H, Kuijer PP, Sluiter JK, Frings-Dresen MH. How to assess physical work-ability with Functional Capacity Evaluation methods in a more specific and efficient way? Work. 2010;37:111-5.

20. Gross DP, Battie MC, Asante AK. Evaluation of a short-form functional capacity evaluation: less may be best. J Occup Rehabil. 2007;17:422-35.

21. Reneman MF, Kuijer W, Brouwer S, Preuper HR, Groothoff JW, Geertzen JH, et al. Symptom increase following a functional capacity evaluation in patients with chronic low back pain: an explorative study of safety. J Occup Rehabil. 2006;16:197-205.

22. Gibson L, Strong J. Safety issues in functional capacity evaluation: findings from a trial of a new approach for evaluating clients with chronic back pain. J Occup Rehabil. 2005;15:237-51.

23. Denier-Bont F, Fischer V, Oesch P, Oliveri M. [Functional Capacity Evaluation: Course manual] ed. Bellikon: Verein IG Ergonomie, Swiss Association of Rehabilitation, 2007.

24. Soer R, Groothoff JW, Geertzen JH, van der Schans CP, Reesink DD, Reneman MF. Pain response of healthy workers following a functional capacity evaluation and implications for clinical interpretation. J Occup Rehabil. 2008;18:290-8.

25. Isernhagen SJ. Functional capacity evaluation: rational, procedure, utility of the kinesiophysical approach. J Occup Rehabil. 1992;2:157-68.

26. Harding VR, Williams AC, Richardson PH, Nicholas MK, Jackson JL, Richardson IH, et al. The development of a battery of measures for assessing physical functioning of chronic pain patients. Pain. 1994;58:367-75.

27. Golding LA. YMCA Fitness Testing and Assessment Manual. 4th ed. Champaign, IL: Human Kinetics, 2000.

28. Pool JJ, Ostelo RW, Hoving JL, Bouter LM, de Vet HC. Minimal clinically important change of the Neck Disability Index and the Numerical Rating Scale for patients with neck pain. Spine (Phila Pa 1976). 2007;32:3047-51.

29. Ferraz MB, Quaresma MR, Aquino LR, Atra E, Tugwell P, Goldsmith CH. Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. J Rheumatol. 1990;17:1022-4.

30. MacDermid JC, Walton DM, Avery S, Blanchard A, Etruw E, McAlpine C, et al. Measurement properties of the neck disability index: a systematic review. J Orthop Sports Phys Ther. 2009;39:400-17.

31. Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale. An updated literature review. J Psychosom Res. 2002;52:69-77.

32. Snaith RP, Zigmond AS. HADS: Hospital Anxiety and Depression Scaleed. Windsor: NFER Nelson, 1994.

33. Borloz S, Trippolini MA, Ballabeni P, Luthi F, Deriaz O. Cross-Cultural Adaptation, Reliability, Internal Consistency and Validation of the Spinal Function Sort (SFS) for French- and German-Speaking Patients with Back Complaints. J Occup Rehabil. 2012;22:387-93.

34. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med. 1990;20:337-40.

35. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307-10.

36. Reneman MF, Soer R, Gerrits EH. Basis for an FCE methodology for patients with work-related upper limb disorders. J Occup Rehabil. 2005;15:353-63.

37. Innes E. Handgrip strength testing: A review of the literature. Aust Occup Ther J. 1999;46:120-40.

38. Reneman MF, Dijkstra PU, Westmaas M, Goeken LN. Test-retest reliability of lifting and carrying in a 2-day functional capacity evaluation. J Occup Rehabil. 2002;12:269-75.

39. Smeets RJ, Hijdra HJ, Kester AD, Hitters MW, Knottnerus JA. The usability of six physical performance tasks in a rehabilitation population with chronic low back pain. Clin Rehabil. 2006;20:989-97.

40. Lygren H, Dragesund T, Joensen J, Ask T, Moe-Nilssen R. Test-retest reliability of the Progressive Isoinertial Lifting Evaluation (PILE). Spine (Phila Pa 1976). 2005;30:1070-4.

41. Altman DG. Some common prolems in medical research. In Altman DG ed. Practical statistics for medical research. London: Chapman & Hall, 1991:396-403.

42. Keller A, Hellesnes J, Brox JI. Reliability of the isokinetic trunk extensor test, Biering-Sorensen test, and Astrand bicycle test: assessment of intraclass correlation coefficient and critical difference in patients with chronic low back pain and healthy individuals. Spine (Phila Pa 1976). 2001;26:771-7.

43. Gross DP, Battie MC. Reliability of safe maximum lifting determinations of a functional capacity evaluation. Phys Ther. 2002;82:364-71.

44. Portney LG, Watkins MP. Reliability. Foundations of clinical research. Applications to practice. 2nd ed. Upper Saddle River, NJ: Prentice-Hall Health, 2000:67.

45. Lomond KV, Cote JN. Shoulder functional assessments in persons with chronic neck/shoulder pain and healthy subjects: Reliability and effects of movement repetition. Work. 2011;38:169-80.

46. Population size and population composition [Federal Statistical Office of Switzerland], 2011. Available from: http://www.bfs.admin.ch/bfs/portal/en/index/themen/01/02.html. Accessed 10.12.2011.

47. Burrus C, Ballabeni P, Deriaz O, Gobelet C, Luthi F. Predictors of nonresponse in a questionnaire-based outcome study of vocational rehabilitation patients. Arch Phys Med Rehabil. 2009;90:1499-505.

48. Isernhagen SJ, Hart DL, Matheson LM. Reliability of independent observer judgments of level of lift effort in a kinesiophysical Functional Capacity Evaluation. Work. 1999;12:145-50.

49. Carroll LJ, Holm LW, Hogg-Johnson S, Cote P, Cassidy JD, Haldeman S, et al. Course and prognostic factors for neck pain in whiplash-associated disorders (WAD): results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. J Manipulative Physiol Ther. 2009;32:S97-S107.

Chapter 5

Construct validity of functional capacity evaluation in patients with whiplash-associated disorders

Maurizio A. Trippolini
Pieter U. Dijkstra
Jan H. B. Geertzen
Michiel F. Reneman

Journal of Occupational Rehabilitation
(Accepted, pending minor revision).

ABSTRACT

Purpose: The construct validity of functional capacity evaluations (FCE) in whiplash-associated disorders (WAD) is unknown. The aim of this study was to analyse the validity of FCE in patients with WAD with cultural differences within a workers' compensation setting.

Method: In a cross-sectional study, 314 participants (42% females, mean age 36.7 years) with WAD (Grade I and II) were referred for an interdisciplinary assessment that included FCE tests. Four FCE tests (hand grip strength, lifting waist to overhead, overhead working, and repetitive reaching) and a number of concurrent variables such as self-reported pain, capacity, disability, and psychological distress were measured. To test construct validity, 29 hypotheses concerning FCE and gender (n=4), and FCE and construct variables (n=20), and FCE and two groups with cultural differences (n=5, 4 for FCE results, 1 for correlations) were formulated a priori and tested.

Results: Men had significantly greater hand grip strength (+17.5 kg) and lifted more weight (+3.7 kg). Regarding the gender-related hypotheses two out of four were not rejected. Hypotheses regarding FCE and construct measures were not rejected in 16 of 20 hypotheses: correlation between FCE and pain ranged from -0.39 to 0.31; FCE and self-reported capacity from -0.42 to 0.61; FCE and disability from -0.45 to 0.34; FCE and anxiety from -0.36 to 0.27; and FCE and depression from -0.41 to 0.34. Five hypotheses regarding FCE and cultural differences were not rejected: 4 hypotheses were not rejected because FCE test results between the two groups differed significantly; 1 hypothesis was not rejected because ES between correlations were small. In total 23 (79%) out of 29 hypotheses were not rejected.

Conclusions: The construct validity was confirmed for the majority of FCE tests for testing functional capacity in patients with WAD with cultural differences and in a workers' compensation setting. Additional validation studies in other settings are needed for verification.

INTRODUCTION

The term whiplash-associated disorders (WAD) has been coined for symptoms related to acceleration-deceleration injuries usually associated with motor vehicle accidents [1]. These symptoms include neck pain, headache, arm pain, and other complaints [1]. The aetiology of WAD likely combines physical and psychological factors; nevertheless, the pathophysiology is not understood [2]. Although the prognosis of WAD is generally favourable, with a recovery rate of 40–60% within the first 12 months, a considerable number of individuals with WAD still reports symptoms and disability, 1 year after the injury [3,4]. Delayed recovery of WAD causes a substantial burden for the individual and society due to long-term sickness, absence, and work disability [5].

According to the guidelines of the International Labor Organization, diseased or disabled persons should be assessed comprehensively to avoid an over- or underestimation of safe work (dis)ability [6]. Functional capacity evaluation (FCE) can be one of the tools included in such an assessment. FCE consists of standardised batteries of functional capacity tests that aim to measure the ability to engage in work-related functioning [7]. When discrepancies between FCE outcomes and the physical workload indicate that capacity is not large enough for the required work load, this capacity may be addressed in rehabilitation programmes to reduce these discrepancies [8, 9]. Moreover, FCEs are used to determine fitness-for-work, and may facilitate the return-to-work process or prelude case closure [10,11].

Functional capacity (FC) has been defined as the highest probable level of function that a person may reach in a domain at a given moment in a standardised environment [8]. Functional capacity is a multidimensional, bio-psycho-social construct, which means that FC is the result of biological and psychological abilities, positively or negatively influenced by personal and external (social) factors (e.g., test environment, education, family) [8,9]. No gold standard exists for the measurement of FC, therefore, validity must be determined by means of construct validity. Construct validity is the degree to which a particular measure relates to other measures in a way one would expect, i.e., in accordance with predefined hypotheses about the correlation or differences between the measures [10]. From a biological perspective, within the bio-psychosocial construct of FC, it can be expected that males are stronger than females and score higher on material handling and grip strength tests, and score similar in postural tolerance and repetitive work tests [11,12]. From a psychological viewpoint it can be hypothesized that in patients with WAD, FC correlates with self-reported pain and mental distress to a larger extend than in healthy workers [4,13]. However, the correlation between FC and mental distress is expected to be smaller compared to the correlation between FCE tests and other measures of functional ability and disability [9,14]. Additionally, the socio-cultural context may influence FC due to different cultural representations and expectations

[15]. A study comparing FCE test results of patients with CLBP in three different countries showed substantial differences between the study samples [16]. People from different ethnical backgrounds living in the same country reported musculoskeletal pain differently [17-19]. One can assume that FCE tests may result in differences between groups with different cultural backgrounds. However, this has not yet been studied.

For both, clinician and researcher it is important to know, how other measures are related to FC, in order to understand what is measured by FCE tests. Because clinical decision-making is based on the results of FCE tests, sound clinimetric properties of FCE tests are required [20]. During the past decades, reliability and, to a lesser extent, validity and safety of FCEs have been studied predominantly in patients with chronic low back pain (CLBP) [10,21] and in one study in healthy persons [13]. FCE validity research should also be conducted in other chronic health conditions such as patients suffering from WAD, because clinimetric properties may not be generalisable across health conditions [22] and cultural settings [23]. Many studies on the construct validity of FCE tests did not meet the requested quality criteria such as formulating an a priori hypothesis for the strength of correlation and adequate sample size [9]. Moreover, few FCE tests were able to demonstrate adequate validity in more than one study, and more than one health condition area [24].

Hence, the aim of this study was to analyse the construct validity of the FCE test for a large sample of patients with WAD, from various cultural backgrounds, who did not return to work after injury onset and who received workers' compensation, using a priori defined hypotheses (Text Box A and B) in a cross-sectional design.

METHODS

Subjects and data collection

Subjects from the German-speaking part of Switzerland were referred by occupational physicians or case managers of the worker's compensation insurance for an interdisciplinary rehabilitation assessment at the rehabilitation clinic in Bellikon (Switzerland). Subjects were insured by the Swiss Accident Insurance Fund (SUVA), the largest accident insurance in Switzerland, which covers injuries from occupational and non-occupational accidents for employed and non-employed subjects. Injured subjects receive compensation of up to 80% of the previous salary, medical and vocational assistance up to a maximum of 2 years, and disability pensions caused by an injury.

The reason for being referred to this assessment was that subjects had not regained full working capacity within 6–12 weeks after the initial injury, had surpassed expected injury

healing times, or had plateaued with medical and other rehabilitative interventions. Inclusion criteria were neck pain due to a whiplash-associated injury according the Québec Task Force (QTF) Classification of WAD, grade I (pain, stiffness, or tenderness without physical signs) or grade II (pain, stiffness, or tenderness with reduced range of motion and point tenderness), sufficient language skills to communicate with the assessors in German language and able to fill out questionnaires in German or Serbo-Croatian, Albanian, Italian, or Spanish (representing the largest immigrant groups in Switzerland) [25], aged 18–65 years, and willingness to participate. Exclusion criteria were main musculoskeletal problem not in the head and neck region, co-morbidity that considerably limited function, such as neurological deficits, rheumatoid diseases, fractures, tumours, osteoporosis, severe psychiatric disorders, pregnancy, and severe cardiac hypertension. All participants were asked for participation prior the interdisciplinary assessment. Participants were informed that they would be allowed to withdraw their participation at any time without disclosing reasons and without consequences for their medical care. The study was performed in accordance with the ethical standards of the Declaration of Helsinki and ethical approval for this study was granted by the Medical Ethics Committee of the Canton Aargau (EK AG 2010/055).

Participants' characteristics were recorded prior to the FCE, and included age, gender, body mass index, marital status, education, native language, duration since injury, education, litigation, work capacity, education status, and physical work demands. After the determination of eligibility for inclusion in the study, patients filled out self-reported measures, i.e., questionnaires (30 min) and carried out FCE tests (20 min).

Measurements

The WAD FCE analysed in this study consisted of tests involving activities of the upper extremities and the neck region, hand grip strength (left and right), lifting waist to overhead, overhead work, and repetitive reaching, left to right and right to left (Appendix 5.1). The reliability of all four FCE tests is good to excellent and the tests are safe in WAD [26]. Participants were briefly instructed on how to perform each test. The evaluator first gave a single demonstration of each test. The lifting test was commenced with a light weight. Participants were then asked to perform the test to their maximum ability. The weights lifted were incrementally increased according to a participant's performance, using weights of 2.5 and 5 kilograms. To determine the level of physical effort, testers used observational criteria indicating physical demand [7]. Testing could be terminated for four reasons: the participant stopped because of, for example, pain; the observer deemed testing to have become unsafe based on biomechanical criteria; heart rate exceeded 85% of the age-related

maximum (220 minus age of the participant); or a predefined time limit was reached. If a participant stopped the lifting waist to overhead test before the criteria for maximum level of demand was observed, the highest weight in kilogram that the patient was willing to lift five times was recorded.

Pain intensity was measured with an 11-point numeric rating scale (NRS) ranging from no pain (0) to worst pain (10). The patient was asked to rate his momentary pain ("pain now"), his worst and his mildest pain during the last 7 days ("maximum pain" and "minimum pain", respectively). The NRS is a commonly used scale with proven reliability and validity in patients with neck pain [27].

The Spinal Function Sort (SFS) was used to measure self-reported functional ability to perform work-related tasks and activities of daily life that involve the spine. The SFS contains 50 drawings with simple verbal descriptions of activities of material handling (e.g. lifting a 10 kg milk-crate from eye-level to the floor), postural tolerance (e.g. wash dishes at a sink) and ambulation (e.g. push and pull a shopping cart). Participants rated functional ability for each activity from "unable" (0) to "able" (4). The SFS yields a single rating ranging from 0 to 200, with higher scores indicating higher or better abilities. The scores can be categorised according the work demands as defined by the Dictionary of Occupational Titles (DOT) [28], allowing a comparison with self-reported functional abilities and work demands (sedentary to lifting weights of over 50 kg). Most patients can fill out the SFS in 10-15min.The SFS has a good reliability and high predictive validity for non-return to work in patients with back pain [14,29].

Neck pain-related disability was measured with the Neck Disability Index (NDI). The NDI contains 10 items: pain intensity, personal care, lifting, reading, headaches, concentration, work, driving, sleeping, and recreation. The scale of each item ranges from no disability (0) to total disability (5). The interpretation for the NDI scores is: 0–4 = none; 5–14 = mild; 15–24 = moderate; 25–34 = severe; over 35–50 = complete disability [30]. The German version of the NDI is reliable and valid [31].

The Hospital Anxiety and Depression Scale (HADS) was used to assess the symptom severity of anxiety disorders and depression in non-psychiatric populations. The HADS consists of two scales, one for anxiety and one for depression (A and D scales, respectively). Each scale contains seven items, with each item rated from 0 (best) to 3 (worst). The scale scores are calculated by summing the responses to the items up to a maximum score of 21 points (severe case) per scale. Scale scores of between 8 and 10 identify mild, 11–15 moderate, and 16 or above severe cases of anxiety/depression. Good reliability and validity, and excellent screening properties have been reported for the use of the HADS in the general population and various clinical populations [32].

A priori hypotheses

Construct validation: known groups

Four hypotheses based on known groups are displayed in Text Box 5A [11,12]. These hypotheses were based on the fact that males are stronger than females, and, therefore, males were expected to outperform females in the strength test, but not in other tests [11].

Text Box 5A A priori hypotheses about the relationship between FCE tests and gender

FCE test	Construct validity is not rejected when mean performance:
Lifting waist to overhead (kg)	females < males (difference ≥10%; p≤0.05)
Isometric hand grip strength right (kgF)	females < males (difference ≥10%;; p≤0.05)
Overhead working (sec)	females ≈ males (difference <10%; p>0.05)
Repetitive reaching right (sec)	females ≈ males (difference <10%; p>0.05)

Construct validation: hypothesis testing

Twenty-five hypotheses on the strength of the association of FCE tests and the additional construct variables were formulated a priori. The theoretical basis for the hypotheses is explained in the introduction. Hypotheses were inferred based on previous studies with patients with chronic low back pain: it was expected that WAD FCE correlates to a higher extend with measures of perceived ability and disability than with measures of mental distress or pain [9,14,33]. The strength of the association is expressed in the absolute value of the correlation coefficient. From the twenty-five, twenty hypotheses were tested about the relationship between four FCE tests and five other construct variables (displayed in Text Box B). Five out of twenty-five hypotheses for two groups with different cultural backgrounds were formulated: four hypothesis regarding the differences of FCE test results between the two groups differed significantly and, one hypothesis was formulated that no major differences in these correlation coefficients (effect size [ES] <0.2) exists between the two groups. Definitions of ES for differences between two correlations are as follow: ES≤0.20 (small), 0.20<ES≤0.50 (medium), 0.50<ES≤0.80 (large) [34]. The two groups with different cultural backgrounds were characterized based on the mother language of the participants.

Data analysis

Normal distribution was visually assessed using P-P plots. Floor and ceiling effects were considered to be present if more than 15% of participants achieved the lowest or highest

Text Box 5B A priori hypotheses about the relationship between 4 FCE tests[a] and 5 other construct variables; pain, perceived functional ability, disability, anxiety, depression

Reference test	Construct validity is not rejected when the strength of the relationship of four FCE tests [a] with	r cut-off values
Pain now (NRS)	pain is low or weak	$0.25 < \|r\| < 0.50$
Self-reported functional ability (SFS)	self-reported functional ability is low to moderate	$0.25 < \|r\| \leq 0.70$
Self-reported disability (NDI)	self-reported disability is moderate	$0.50 \leq \|r\| \leq 0.70$
Anxiety (HADS A)	anxiety is low or weak	$0.25 < \|r\| < 0.50$
Depression (HADS D)	depression is low or weak	$0.25 < \|r\| < 0.50$

[a] FCE includes the tests Lifting waist to overhead (kg), Hand grip strength right, (kgF), Overhead working (sec), Repetitive reaching right (sec); $|r|$ = Correlation Coefficient, absolute value. The direction of the association depends on the scoring of the reference measure. Interpretation: 0.00–0.25 little if any ("not correlated"); 0.26–0.49 low or weak; 0.50–0.69 moderate; 0.70–0.89 high or strong; 0.90–1.00 very strong correlation [35].

possible score of the overhead working test [36]. The overhead working test was expected to display ceiling effects because the test was limited to a maximum of 5 min.

Associations were calculated using Pearson correlation coefficient for bivariate normally distributed data, or else a Spearman rank correlation coefficient. For relationships between gender and overhead working, and repetitive reaching, respectively, equivalence testing was performed [37]. Equivalence is established if 10% the margins of differences between gender fall within the 90% confidence intervals of the difference [37]. To analyse differences between genders and between two groups with different cultural backgrounds, independent sample t-test, a Mann-Whitney U Test, χ^2 test, or linear regression was used as appropriate. The validity of the WAD FCE was considered not rejected when no ceiling or floor effects were observed in the FCE tests and the majority (80%) of the 29 a priori hypotheses were not rejected [38]: four hypotheses concerning the relationship between FCE tests and gender, 20 hypotheses concerning the associations of the FCE tests and the other construct variables and five hypotheses concerning the two groups with different cultural backgrounds. Validity was not rejected when, significant differences in FCE test results emerged between the two groups in all 4 comparisons, and the ES for differences in correlations between FCE tests and the five construct variables between both groups was ≤0.2 in 16 or more of the 20 comparisons. The ES for differences between correlations of the two groups were calculated by subtracting the Z score of the German mother language group by the Z score of the non-German mother language group. Z scores were calculated as follows: 0.5 ln [(1+r)/(1-r)], were r is the correlation coefficient between an FCE test and

a reference measure [34]. P<0.05 was used as a cut-off, indicating statistical significance. All analyses were performed using SPSS (Statistical Package for Social Sciences, Version 21, IBM Corp.).

RESULTS

Participants

From January 2011 to January 2012, 428 patients were referred for interdisciplinary assessment due to delayed recovery after musculoskeletal injury. From the referred patients (n=114), 79 (69%) were not eligible because the main problem was not in the neck and head region; 17 (15%) had insufficient German language skills to communicate with the assessors or not able to fill out the questionnaires in the language versions available; 5 (5%) had acute comorbidity that limited testing, such as fracture or severe psychiatric disorder; 2 (2%) were pregnant; 6 (5%) were excluded due to other medical reasons; 3 (3%) due to age under 18 or over 65 years; and 2 (2%) were of grade III–IV by QTF criteria.

In total, 314 patients fulfilled the inclusion criteria and participated in this study. The participants' characteristics are presented in Table 5.1. Participants' characteristics were analysed in two groups with cultural differences, n=152 (48%) participants with German as their mother language and n=162 (52%) with a non-German language as their mother language. Significant differences between the groups were observed in 8 out of 10 main participant characteristics (Table 5.1). In five self-reported measures (Table 5.1), significant differences were found between the two groups.

Descriptive analysis of FCE test results

Normal distribution was found in three out of four FCE tests, i.e., lifting waist to overhead, hand grip strength (right), and repetitive reaching (right). A ceiling effect was observed in the overhead working test with 38% (n=119) of the participants reaching the maximum time limit of 300 sec. Between the two language groups and genders, the differences in FCE tests were significant in six out of eight comparisons (Table 5.2). There was no significant interaction between gender and language.

Construct validation: known groups

As presented in Table 5.3, men had a significantly greater hand grip strength (+17.5 kg), and lifted significantly more weight over head (+3.7 kg). Differences between genders were in

Table 5.1 Characteristics of the participants

Characteristics, unit or scale	Total n=314	German n=152	Non-German* n=162	P-value [e]
Age (years), Median (IQR)**	36.0 (27.0–45.0)	34.5 (26.0–46.0)	36.0 (29.9–44.3)	<.476 [f]
Gender female, n (%)	133 (42.4)	83 (54.6)	50 (30.9)	<.001 [h]
BMI [a], Median (IQR)**	26.0 (22.0–30.0)	24.0 (21.0–29.0)	27.0 (24.0–30.0)	<.001 [f]
Marital status, n (%)				
Married or co-habitation	161 (51.3)	40 (26.3)	121 (74.1)	<.001 [g]
Single	109 (34.7)	85 (55.9)	24 (14.8)	
Divorced or living separated	42 (13.4)	26 (17.1)	16 (9.9)	
Other	2 (0.6)	1 (0.7)	1 (0.6)	
Duration since WAD injury claim opening (days), Median (IQR)	91.0 (72–124.0)	91.0 (72.0–122.5)	91.0 (73.5–126.3)	<.986 [f]
Attorney involved, n (%)	86 (27.4)	37 (24.3)	49 (30.2)	<.025 [g]
Work incapacity in % previous work [b], Median (IQR)	80 (40–100)	50 (25–100)	100 (50–100)	<.001 [f]
Education [c], n (%)				
Low	147 (46.8)	33 (21.8)	114 (70.4)	<.001 [g]
Intermediate	159 (50.6)	113 (74.3)	46 (28.4)	
High	8 (2.5)	6 (3.9)	2 (1.2)	
Physical work demands [d] n (%)				
Sedentary to light (<5–10 kg)	110 (35.0)	74 (48.7)	36 (22.2)	<.001 [g]
Light to medium (11–25 kg)	113 (36.0)	42 (27.7)	71 (43.8)	
Heavy to very heavy (26 to >45 kg)	91 (29.0)	36 (23.6)	55 (34.0)	
Pain intensity (NRS 0–10) Mean (SD)				
Pain now Mean (SD)	4.6 (2.2)	4.2 (2.3)	4.9 (2.2)	<.002 [i]
Pain maximum, last 7 days, Median (IQR)**	8.0 (6.0–9.0)	7.5 (5.3–8.0)	8.0 (6.8–9.0)	<.011 [f]
Pain minimum, last 7 days, Median (IQR)**	3.0 (1.0–4.0)	2.0 (0.0–3.0)	3.0 (2.0–5.0)	<.001 [f]
Perceived functional ability (SFS 0–200), Median (IQR)** ,***	141.0 (103–163)	151.7 (128–174)	120.0 (91–158)	<.001 [f]
Disability (NDI 0–50) Mean (SD	22.5 (8.3)	20.9 (7.9)	24.0 (8.3)	<.001 [i]

Table 5.1 continues on next page

Table 5.1 *Continued*

Characteristics, unit or scale	Total n=314	German n=152	Non-German* n=162	P-value [e]
Anxiety (HADS 0–21), Median (IQR)**	9.0 (5.0–12.0)	6.0 (4.0–10.0)	11.0 (7.0–14.0)	<.001 [f]
Depression (HADS 0–21), Median (IQR)**	7.0 (3.8–10.0)	5.0 (2.0–8.0)	8.5 (5.8–12.00)	<.001 [f]

* Albanian n=82 (62.1%), Serbo-Croatian n=25 (8%), Italian=17 (5.5%), Other n=28 (8.8%; Turkish, Arabic, Portuguese, Spanish). **Data with a skewed distribution are presented with a median and an interquartile range (IQR). *** Data missing for 7 participants. [a] BMI = Body mass index; [b] work incapacity set by the insurance assessed for the actual or previous job (if jobless) in % at the time of WAD FCE; [c] low = no vocational education, intermediate = vocational education, high = bachelor or higher education; [d] Maximum physical work load of material handling tasks according to the Dictionary of Occupational Titles (DOT). Category light to medium was added to ensure that all participants could be categorized in a continuous scale. NRS = Numeric rating scale; NDI = Neck Disability Questionnaire; HADS = Hospital Anxiety Depression Scale; SFS = Spinal Function Sort. [e] p-value = significant, if $p<0.05$ concerning differences between men and female based on the results of [f] Mann-Whitney U Test, [g] skewed distribution of scaled data, [h] χ^2-test for categorical data, and [i] t-test for continuous data.

the overhead working test -7.4 sec and the repetitive reaching test -8.2 sec. The 10% margin of differences between gender for overhead working was 18.5 sec. (90% CI -26.2 to 11.4) and for repetitive reaching 8.8 sec (90% CI 3.2 to 13.2). The 90% CI did not fall within the 10% margin, thus non equivalence could not be ruled out. Two out of four gender-related hypotheses were not rejected.

Construct validation: hypothesis testing

Correlations between the FCE tests and pain, perceived functional ability, disability, anxiety, and depression are presented in Table 5.4. For each of the FCE tests, four out of five hypotheses were not rejected.

Correlations for the two language groups between the four FCE tests and the reference measures are presented in Table 5.5. Eighteen out of 20 ES were ≤0.20 (ranging from 0.01 to 0.16). In two comparisons, the ES for the difference in correlations between groups with different cultural backgrounds was >0.20; -0.21 for lifting waist to overhead and the SFS, and 0.22 for lifting waist to overhead and HADS anxiety (ES data available from the author on request). The hypothesis on the validity of FCE tests in patients with cultural differences was not rejected because ES were ≤0.20 in the 18 of 20 comparisons.

Table 5.2 Differences in FCE results between language groups and gender

	German		Non-German		P-value*	
FCE tests (unit), Mean (SD)	Males n=69	Females n=83	Males n=112	Females n=50 [a]	Gender differences	Language differences
Hand grip strength right (kgF)	45.9 (12.1)	26.0 (8.1)	37.3(12.9)	18.4 (8.2)	<.001	<.001
Lifting waist to overhead (kg)	14.8 (6.4)	10.3 (4.0)	11.9 (6.0)	7.3 (3.7)	<.001	<.001
Overhead working (sec)	228.2 (90.0)	222.3 (94.9)	157.8 (95.9)	141.4 (92.0)	.322	<.001
Repetitive reaching right (sec) [a]	76.9 (20.3)	70.7 (25.2)	88.4 (28.1)	84.63 (28.8)	.098	<.001

SD = standard deviation; [a] data missing for one participant; * based on the results of a linear regression analysis.

Table 5.3 Differences in FCE tests results between genders

		Males n=181		Females n=133			
Hypotheses	FCE tests (unit)	Mean	SD	Mean	SD	P-value [a]	Interpretation of hypothesis
1	Hand grip strength right (kgF)	40.6	13.3	23.1	8.9	<.001	not rejected
2	Lifting waist to overhead (kg)	13.0	6.3	9.2	4.1	<.001	not rejected
3	Overhead working (sec)	184.6	99.4	192.0	101.4	.557 #	rejected [b]
4	Repetitive reaching right (sec)	84.0	26.0	75.8	27.3	<.001 #	rejected [b]

SD = standard deviation; "ceiling effect" at 300 sec.; [a] p-value = significant, if $p<0.05$; # Mann-Whitney U Test; [b] hyptheses rejected, based on results of equivalence testing.

DISCUSSION

The aim of the study was to analyse construct validity of FCE tests for application in patients on workers' compensation due to WAD across groups with cultural differences (defined as the mother language of the patient 22 Out 29 (79%) instead of the expected 80% of the a priori defined hypotheses were not rejected. Not rejected were 2 out of 4 gender-related hypotheses, 5 out of 5 culture-related hypotheses, and 16 out of 20 construct-related hypotheses. Differences in correlations between the groups with cultural differences were statistically significant, but small (18 out of 20 ES were ≤0.2), despite large differences in patient characteristics and FCE

Table 5.4 Correlations between the results of FCE tests and pain, perceived functional ability, disability, anxiety, and depression to test construct validity of FCE tests, for the total group

FCE tests	Pain now (NRS)	Functional ability (SFS)	Disability (NDI)	Anxiety (HADS A)	Depression (HADS D)	Total of hypotheses not rejected
Hand grip strength right (kgF) 95% CI	-0.26 (-0.36 to -0.16)	0.38 (0.28 to 0.47)	-0.26 (-0.36 to -0.15)	-0.28 (-0.38 to -0.17)	-0.25 (-0.35 to -0.15)	4 out of 5
Lifting waist to overhead (kg) 95% CI	-0.39 (-0.48 to -0.29)	0.60 (0.52 to 0.66)	-0.39 (-0.48 to -0.29)	-0.27 (-0.37 to -0.16)	-0.30 (-0.40 to -0.20)	4 out of 5
Overhead working (sec) 95% CI	-0.36 (-0.46 to -0.26)	0.61 (0.54 to 0.68)	-0.45 (-0.53 to -0.35)	-0.36 (-0.45 to -0.26)	-0.41 (-0.50 to -0.31)	4 out of 5
Repetitive reaching right (sec) 95% CI	0.31 (0.20 to 0.40)	-0.42 (-0.50 to -0.32)	0.34 (0.23 to 0.43)	0.27 (0.16 to 0.37)	0.34 (0.24 to 0.43)	4 out of 5
Number of hypotheses not rejected	4 out of 4	4 out of 4	0 out of 4	4 out of 4	4 out of 4	16 out of 20
Interpretation of hypothesis	Hypotheses not rejected	Hypotheses not rejected	Hypotheses rejected	Hypotheses not rejected	Hypotheses not rejected	Construct validity not rejected

The Pearson correlation statistic was used. All correlations were significant at the p-value 0.01 level (2-tailed). Interpretation: NRS = Numeric Rating Scales; SFS = Spinal Function Sort; NDI = Neck Disability Questionnaire; HADS = Hospital Anxiety Depression Scale. CI = Confidence interval.

Table 5.5 Overview of correlations overview between the results of FCE tests and pain, perceived functional ability, disability, anxiety, and depression separated by language groups

FCE tests	Pain now (NRS 0–10)		Functional ability (SFS 0–200)		Disability (NDI 0–50)		Anxiety (HADS A 0–21)		Depression (HADS D 0–21)	
	German	N-German	German	N-German	German	N-German	German	N-German	German	N-German
Hand grip strength right (kgF) [a] 95% CI	-0.24* -0.38 to -0.08	-0.26 -0.40 to -0.11	*0.26* *0.10 to 0.40*	*0.43* *0.30 to 0.55*	-0.16* -0.31 to 0.00	-0.32 -0.45 to -0.17	-0.24 -0.38 to -0.08	-0.27 -0.40 to -0.12	-0.18* -0.33 to -0.22	-0.27 -0.40 to -0.12
Lifting waist to overhead (kg) [a] 95% CI	-0.41 -0.53 to -0.26	-0.33 -0.46 to -0.19	*0.50*[c] *0.37 to 0.61*	*0.64*[c] *0.54 to 0.73*	-0.34 -0.47 to -0.19	-0.40 -0.52 to -0.26	*-0.12*†[d] *-0.27 to 0.04*	*-0.32*[d] *-0.45 to -0.18*	-0.19* -0.34 to -0.03	-0.32 -0.45 to -0.18
Overhead working (sec) [b] 95% CI	-0.39 -0.52 to -0.25	-0.28 -0.41 to -0.13	0.60 0.48 to 0.69	0.52 0.40 to 0.63	-0.42 -0.54 to -0.28	-0.40 -0.52 to -0.26	-0.20* -0.35 to -0.04	-0.30 -0.43 to -0.15	-0.26 -0.41 to -0.11	-0.35 -0.48 to -0.20
Repetitive reaching right (sec) [b] 95% CI	0.28 0.13 to 0.42	0.27 0.12 to 0.41	-0.42 -0.54 to -0.28	-0.36* -0.49 to -0.21	0.39 0.25 to 0.52	0.29 0.14 to 0.43	0.17* -0.41 to -0.11	0.19* 0.04 to 0.34	0.30 0.14 to 0.44	0.26 0.11 to 0.40

[a] Pearson correlation statistic, [b] Spearman's rank correlation statistic. All values presented were significant at $p<0.01$ (2-tailed), except * $p<0.05$ (2-tailed), † $p=0.15$. Non-italics: difference in correlation between cultural groups ES≤0.20. *Italics*: difference ES >0.20: [c] ES = 0.214, [d] ES = 0.215. NRS = Numeric Rating Scales; SFS = Spinal Function Sort; NDI = Neck Disability Questionnaire; HADS = Hospital Anxiety Depression Scale. CI = Confidence interval.

performances. A ceiling effect was observed in 1 test (overhead working). Overall, the construct validity was confirmed for the majority of FCE tests for testing functional capacity in patients with WAD with cultural differences and in a workers' compensation setting.

The results of the study support the bio-psycho-social construct of FCE in WAD: we observed differences between males and females (bio), between language groups (socio), and small but consistent relationships with psychological factors (psycho). The gender differences in FCE tests in this study are consistent with the results of others [11]. Differences in test results, but not in correlations, were observed between language groups. The non-German language group consisted of individuals from the largest immigrant groups in Switzerland [25]. The participants of this study consisted of 52% whose mother language was non-German, which is higher than the 18% of the Swiss population [25]. The proportion of male participants in the non-German group in this study was similar (47.6%) to that of the Swiss working population (51%) [25], but higher than usually reported in WAD [1]. These differences may be explained by the fact the study participants were insured by SUVA, which insures many companies from the industry and construction sector, where the rate of male, non-German speaking subjects is higher than in the other business sectors [39]. Many immigrants have been naturalised to Swiss citizenship, hence mother language was chosen as an indicator for cultural differences. Mother language has been reported as a valid indicator for cultural differences [40]. A study on the coping styles of patients with low back pain found large differences among groups with different mother languages in Switzerland [41].

To test construct validity, associations were made with other constructs known to be associated with FCE outcomes. In two out of four instances, the associations between gender and FCE outcomes occurred as hypothesized. Although differences were small in the overhead working and repetitive working tests, equivalence between genders could not be ruled out. We expected no difference between genders, because for this test muscle force is not likely primary factor for outcome. In the healthy population, conflicting evidence for the difference between genders in dexterity performance tests has been reported [12,42,43]. Results in fine manual dexterity tests may be influenced by finger size; smaller fingers were related to better outcomes [44]. This might be a plausible explanation of the results of this study.

In patients with CLBP, moderate correlations between FCE and SFS [14], and between FCE and other self-reported measures of disability were reported [9]. In this study, FCE correlated more strongly with SFS (moderate correlations) than with the NDI (weak correlations). There could be several explanations for this. Firstly, the items of the SFS more closely resemble the items of the FCE than the NDI. Secondly, inconsistent wording of the NDI items concerning the influence of pain on activity levels may partly explain the results. Thirdly, while our hypothesis was based on the majority of the studies in CLBP where the relationship between

FCE and self-reported disability was moderate, this relationship may be slightly different in patients with WAD or when using the NDI. Additionally, there may have been unknown sample characteristics contributing to these differences.

The strengths of the correlations between FCE and psychological variables in patients with WAD appear higher compared with CLBP patients [9]. This may be consistent with the relevance of psychological factors in WAD [3,45]. We compared our results with a recently published study with 40 patients with WAD from the Netherlands [46]. On average the Dutch sample was younger (Mean 33 years, SD 9.6), more female (55%) and the duration since whiplash injury was longer (median 12 months, IQR 7–19). While the results of the repetitive reaching test between the two samples were similar (mean difference 2 seconds), the differences between the lifted weight from waist to overhead between the Dutch and the Swiss patients with WAD was substantial (the Dutch lifted a mean of 12.2 kg more). The differences between the studies might be explained by sample variation since sample in the Dutch study was small. But these differences need further investigation. Nevertheless, they are consistent with a study that reported large differences in FCE outcomes between different countries in patients with low back pain [19]. The strength of the correlations between NDI and lifting waist to overhead and overhead working between the Dutch and the Swiss WAD samples were similar, suggesting some robustness of the results between study samples from different countries. Shortly, these findings underline the importance of replication of validation studies among different (social security) contexts.

Some potential limitations have to be addressed. The study population consisted of injured workers who did not return to work within the first 6 to 12 weeks, for whom recovery had plateaued, and who were referred by the case manager or occupational physician. The validity of WAD FCE should also be established in other WAD patients outside the workers' compensation setting, in general practice or in more chronic WAD patients (in rehabilitation settings). Moreover, the a priori defined hypotheses were based on previous studies performed in populations other than WAD. Most studies reported conflicting evidence on many FCE-related factors [9], so cut-offs for the strength of the correlation were arbitrarily chosen. Additionally, if other measures for construct validation had been used, the results might have been different. In this study, self-reported measures were used, which are related to physical capacity but distinct [47-49].

In the overhead working test, a ceiling effect was found in 38% of the participants, as reported for healthy subjects and CLBP patients [50,51]. It was not expected that such a high proportion of patients with WAD would reach the time limit of 300 sec, because one could suppose a reduced postural tolerance in the neck and upper limbs. For future research, we suggest modifying the overhead working test by having the subject wear two cuff weights of 1 kg

each around on their forearm to reduce ceiling effects, as described for healthy subjects [52].

The strengths of this validation study of FCE for WAD patients were the use of a priori defined hypotheses in the analyses, allowing transparency and explicitness. Therefore, several comparisons could be made to a variety of constructs, enabling the reader to interpret the validity from different points of views. Additionally, the design and the sample size of the current study meet the proposed quality standards for FCE validation studies [22]. Moreover, patients with different cultural backgrounds participated in our study, unless previous FCE studies where languages or cultural differences were not reported [9]. To our knowledge, this has not been the subject of a study in a setting similar to ours (validation of FCE tests). Although replication is needed, the results of this study support the validity of the WAD FCE in patients with different languages as their mother language (i.e., cultural backgrounds).

CONCLUSION

The construct validity was confirmed for the majority of FCE tests for testing functional capacity in patients with WAD with cultural differences and in a workers' compensation setting. Additional validation studies in other settings are needed for verification.

ACKNOWLEDGEMENTS

The authors thank the physiotherapists and physicians of the Department of Work Rehabilitation, Rehaklinik Bellikon, for their help in performing the tests and collecting data. We also thank Roy Stewart for statistical advice, Peter Erhart, Claudia Diethelm, and Axel Gehrke for data preparation, technical and administrative support, and all patients for their participation.

REFERENCES

1. Spitzer WO, Skovron ML, Salmi LR, Cassidy JD, Duranceau J, Suissa S, et al. Scientific monograph of the Quebec Task Force on Whiplash-Associated Disorders: redefining "whiplash" and its management. Spine (Phila Pa 1976). 1995;20:1S-73S.
2. Curatolo M, Bogduk N, Ivancic PC, McLean SA, Siegmund GP, Winkelstein BA. The role of tissue damage in whiplash-associated disorders: discussion paper 1. Spine (Phila Pa 1976). 2011;36:S309-15.
3. Kamper SJ, Rebbeck TJ, Maher CG, McAuley JH, Sterling M. Course and prognostic factors of whiplash: a systematic review and meta-analysis. Pain. 2008;138:617-29.

4. Carroll LJ, Hogg-Johnson S, Cote P, van der Velde G, Holm LW, Carragee EJ, et al. Course and prognostic factors for neck pain in workers: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine (Phila Pa 1976). 2008;33:S93-100.

5. Chappuis G, Soltermann B. Number and cost of claims linked to minor cervical trauma in Europe: results from the comparative study by CEA, AREDOC and CEREDOC. Eur Spine J. 2008;17:1350-7.

6. ILO. Technical and ethical guidelines for workers' health surveillance. Occupational Safety and Health Series No. 72. International Labor Office [International Labor Organisation]. Geneva, 1998.

7. Isernhagen SJ. Functional capacity evaluation: rational, procedure, utility of the kinesiophysical approach. J Occup Rehabil. 1992;2:157-68.

8. Soer R, van der Schans CP, Groothoff JW, Geertzen JH, Reneman MF. Towards consensus in operational definitions in functional capacity evaluation: a Delphi Survey. J Occup Rehabil. 2008;18:389-400.

9. van Abbema R, Lakke SE, Reneman MF, van der Schans CP, van Haastert CJ, Geertzen JH, et al. Factors associated with functional capacity test results in patients with non-specific chronic low back pain: a systematic review. J Occup Rehabil. 2011;21:455-73.

10. Innes E. Reliability and validity of functional capacity evaluations: an update. International Journal of Disability Management Research. 2006;1:135-48.

11. Soer R, van der Schans CP, Geertzen JH, Groothoff JW, Brouwer S, Dijkstra PU, et al. Normative values for a functional capacity evaluation. Arch Phys Med Rehabil. 2009;90:1785-94.

12. Haward BM, Griffin MJ. Repeatability of grip strength and dexterity tests and the effects of age and gender. Int Arch Occup Environ Health. 2002;75:111-9.

13. Lakke SE, Soer R, Geertzen JH, Wittink H, Douma RK, van der Schans CP, et al. Construct validity of functional capacity tests in healthy workers. BMC Musculoskelet Disord. 2013;14:180.

14. Oesch PR, Hilfiker R, Kool JP, Bachmann S, Hagen KB. Perceived functional ability assessed with the spinal function sort: is it valid for European rehabilitation settings in patients with non-specific non-acute low back pain? Eur Spine J. 2010;19:1527-33.

15. Sloots M, Dekker JH, Pont M, Bartels EA, Geertzen JH, Dekker J. Reasons of drop-out from rehabilitation in patients of Turkish and Moroccan origin with chronic low back pain in The Netherlands: a qualitative study. J Rehabil Med. 2010;42:566-73.

16. Reneman MF, Kool J, Oesch P, Geertzen JH, Battie MC, Gross DP. Material handling performance of patients with chronic low back pain during functional capacity evaluation: a comparison between three countries. Disabil Rehabil. 2006;28:1143-9.

17. Palmer B, Macfarlane G, Afzal C, Esmail A, Silman A, Lunt M. Acculturation and the prevalence of pain amongst South Asian minority ethnic groups in the UK. Rheumatology (Oxford). 2007;46:1009-14.

18. Allison TR, Symmons DP, Brammah T, Haynes P, Rogers A, Roxby M, et al. Musculoskeletal pain is more generalised among people from ethnic minorities than among white people in Greater Manchester. Ann Rheum Dis. 2002;61:151-6.

19. Scheermesser M, Bachmann S, Schamann A, Oesch P, Kool J. A qualitative study on the role of cultural background in patients' perspectives on rehabilitation. BMC Musculoskelet Disord. 2012;13:5.

20. King PM, Tuckwell N, Barrett TE. A critical review of functional capacity evaluations. Phys Ther. 1998;78:852-66.

21. Schiphorst Preuper HR, Reneman MF, Boonstra AM, Dijkstra PU, Versteegen GJ, Geertzen JH, et al. Relationship between psychological factors and performance-based and self-reported disability in chronic low back pain. Eur Spine J. 2008;17:1448-56.

22. Reneman M, Wittink H, Gross DP. The scientific status of functional capacity evaluation. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations: American Medical Association, 2009:393-420.

23. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1976). 2000;25:3186-91.

24. Innes E, Straker L. Validity of work-related assessments. Work. 1999;13:125-52.

25. Population size and population composition [Swiss Federal Statistical Office], 2013. Available from: http://www.bfs.admin.ch/bfs/portal/de/index/themen/01/02.html. Accessed 23.14.2013.

26. Trippolini MA, Reneman MF, Jansen B, Dijkstra PU, Geertzen JH. Reliability and safety of functional capacity evaluation in patients with whiplash associated disorders. J Occup Rehabil. 2013;23:381-90.

27. Pool JJ, Ostelo RW, Hoving JL, Bouter LM, de Vet HC. Minimal clinically important change of the Neck Disability Index and the Numerical Rating Scale for patients with neck pain. Spine (Phila Pa 1976). 2007;32:3047-51.

28. U.S. Department of Labor. The Revised Handbook for Analyzing Jobs. 4th ed. Indianapolis: JIST Works, inc., 1991.

29. Borloz S, Trippolini MA, Ballabeni P, Luthi F, Deriaz O. Cross-Cultural Adaptation, Reliability, Internal Consistency and Validation of the Spinal Function Sort (SFS) for French- and German-Speaking Patients with Back Complaints. J Occup Rehabil. 2012;22:387-93.

30. Vernon H. The Neck Disability Index: state-of-the-art, 1991-2008. J Manipulative Physiol Ther. 2008;31:491-502.

31. Swanenburg J, Humphreys K, Langenfeld A, Brunner F, Wirth B. Validity and reliability of a German version of the Neck Disability Index (NDI-G). Man Ther. 2014;19:52-8.

32. Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale. An updated literature review. J Psychosom Res. 2002;52:69-77.

33. Smeets RJ, van Geel AC, Kester AD, Knottnerus JA. Physical capacity tasks in chronic low back pain: what is the contributing role of cardiovascular capacity, pain and psychological factors? Disabil Rehabil. 2007;29:577-86.

34. Hojat M, Xu G. A visitor's guide to effect sizes: statistical significance versus practical (clinical) importance of research findings. Adv Health Sci Educ Theory Pract. 2004;9:241-9.

35. Hazard Munro B. Statistical Methods for Health Care. Philadelphia: J. B. Lippincott, 1986.

36. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res. 1995;4:293-307.

37. Walker E, Nowacki AS. Understanding equivalence and noninferiority testing. J Gen Intern Med. 2010;26:192-6.

38. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60:34-42.

39. Suva. Suva: an overview [Swiss Accident Insurance Fund] 2013. Available from: http://www.suva.ch/english/startseite-en-suva/suva-en-suva/ueberblick-en-suva.htm. Accessed 17.09.2013.

40. Burrus C, Ballabeni P, Deriaz O, Gobelet C, Luthi F. Predictors of nonresponse in a questionnaire-based outcome study of vocational rehabilitation patients. Arch Phys Med Rehabil. 2009;90:1499-505.

41. Schulz PJ, Hartung U, Riva S. Causes, coping, and culture: a comparative survey study on representation of back pain in three swiss language regions. PLoS One. 2013;8:e78029.

42. Amirjani N, Ashworth NL, Gordon T, Edwards DC, Chan KM. Normative values and the effects of age, gender, and handedness on the Moberg Pick-Up Test. Muscle Nerve. 2007;35:788-92.

43. Jimenez-Jimenez FJ, Calleja M, Alonso-Navarro H, Rubio L, Navacerrada F, Pilo-de-la-Fuente B, et al. Influence of age and gender in motor performance in healthy subjects. J Neurol Sci. 2011;302:72-80.

44. Peters M, Servos P, Day R. Marked sex differences on a fine motor skill task disappear when finger size is used as covariate. J Appl Psychol. 1990;75:87-90.

45. Carroll LJ, Holm LW, Hogg-Johnson S, Cote P, Cassidy JD, Haldeman S, et al. Course and prognostic factors for neck pain in whiplash-associated disorders (WAD): results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine (Phila Pa 1976). 2008;33:S83-92.

46. van der Meer S, Reneman MF, Verhoeven J, van der Palen J. Relationship between self-reported disability and functional capacity in patients with Whiplash Associated Disorder. J Occup Rehabil. 2013. DOI 10.1007/s10926-013-9473-6.

47. Wittink H, Rogers W, Sukiennik A, Carr DB. Physical functioning: self-report and performance measures are related but distinct. Spine (Phila Pa 1976). 2003;28:2407-13.

48. Reneman MF, Jorritsma W, Schellekens JM, Goeken LN. Concurrent validity of questionnaire and performance-based disability measurements in patients with chronic nonspecific low back pain. J Occup Rehabil. 2002;12:119-29.

49. Lin CW, McAuley JH, Macedo L, Barnett DC, Smeets RJ, Verbunt JA. Relationship between physical activity and disability in low back pain: a systematic review and meta-analysis. Pain. 2011;152:607-13.

50. Reneman MF, Brouwer S, Meinema A, Dijkstra PU, Geertzen JH, Groothoff JW. Test-retest reliability of the Isernhagen Work Systems Functional Capacity Evaluation in healthy adults. J Occup Rehabil. 2004;14:295-305.

51. Brouwer S, Reneman MF, Dijkstra PU, Groothoff JW, Schellekens JM, Goeken LN. Test-retest reliability of the Isernhagen Work Systems Functional Capacity Evaluation in patients with chronic low back pain. J Occup Rehabil. 2003;13:207-18.

52. Soer R, Gerrits EH, Reneman MF. Test-retest reliability of a WRULD functional capacity evaluation in healthy adults. Work. 2006;26:273-80.

APPENDIX: MATERIALS AND PROCEDURES OF THE WAD FCE

Isometric hand grip strength

Isometric hand grip strength was measured in a seated position. The subjects held their shoulder adducted without internal or external rotation, elbow flexed at approximately 90° and the forearm and wrist in neutral position. Grip strength of the right and left hand was measured in a three-trial procedure while maintaining in a hand dynamometer in one single handgrip position adapted to the handsize of the subject (Jamar PC 5030, Preston Corporation, 1994). An average amount of kgF was scored.

Lifting waist to overhead test

Lifting waist to overhead was measured during 5 lifts of the crate from table to crown in standing position, and vice versa within 90 sec in standing position. The test was executed with a wooden crate (40 x 30 x 26 cm) of 2.5 kg. Weight increments of 2.5 kg or 5 kg each were used until the maximum amount of weight was reached. Maximum performance was recorded in kg.

Overhead work test

Overhead working was performed standing with hands at crown height for manipulation of nuts and bolts. The ceiling of the test was 5 minutes. The time that the position was held was recorded (sec).

Repetitive reaching test

Repetitive reaching was determined by fast horizontal movements of the upper extremity in a sitting position. Marbles were removed from bowls at arm length distance at table height from left to right and vice versa, with right and then left arm. The time taken to remove 30 marbles was recorded (sec).

Chapter 6

Can functional capacity tests predict future work capacity in patients with whiplash-associated disorders?

Maurizio A. Trippolini
Pieter U. Dijkstra
Pierre Côté
Stefan M. Scholz-Odermatt
Jan H. B. Geertzen
Michiel F. Reneman

Archives of Physical Medicine and Rehabilitation
(Accepted, contigent on some revisions).

ABSTRACT

Objective: To determine whether Functional Capacity Evaluation (FCE) tests predict future work capacity of patients with whiplash-associated disorders (WAD).

Design: Prospective cohort study.

Setting: Rehabilitation center.

Participants: Sick listed workers with WAD 6 to 12 weeks after injury.

Interventions: Patients performed 8 work-related FCE tests.

Main outcome measures: Work capacity (WC; 0–100%) measured at baseline and 1, 3, 6, and 12 months after testing. Correlation coefficients between FCE tests and WC were calculated. A linear mixed model analysis was used to assess the association between FCE and future WC.

Results: In total 267 patients with WAD grade I or II participated in the study. Mean WC increased over time from 20.8 (SD 27.6) at baseline to 32.3 (SD 38.4), 51.3 (SD 42.8), 65.6 (SD 42.2), 83.2 (SD 35.0) at 1, 3, 6, 12 months follow-up respectively. Correlation coefficients between FCE tests and WC ranged between r=0.06 (lifting low at 12 months follow-up) to r=0.39 (walking speed at 3 months). Strength of correlations decreased over time. FCE tests did not predict WC at follow-up. The predictors of WC were ln (time) (β = 23.74), mother language (β=5.49), work capacity at baseline (β=1.01), and self-reported disability (β=-0.20). Two interaction terms ln (time) x workcapacity (β=-0.19), and ln (time) x self-reported disability (β=-0.21) were significant predictors of WC.

Conclusion: FCE tests performed within 3 months after WAD injury are associated to WC at baseline, but do not predict future WC, whereas time course, mother language, WC at baseline, and self-reported disability do predict future WC. Additionally, interaction between time course WC at baseline and self-reported disability respectively predicted future WC.

INTRODUCTION

The prognosis of whiplash-associated disorders (WAD) is generally favorable, with recovery rates of 40–60% within the first year. However, many individuals with WAD reports symptoms and disability one year after the injury [1,2]. Delayed recovery from WAD causes a substantial burden to individuals and society caused by long term sickness absence and work disability [3]. Several studies have investigated prognostic factors for the clinical course of WAD [1,4]. Established prognostic factors include post-injury pain intensity and self-reported disability [5]. Psychosocial factors such as fear of movement, self-efficacy beliefs, poor recovery expectation, pain catastrophizing, passive coping and depression predict poor recovery [1,4,6,7]. Studying the prognosis of whiplash is complicated and the validity of previous studies has been limited by small sample size, inclusion of patients >6 months after injury onset, short follow-up periods (<6 months), loss to follow-up, unblinded outcome assessors and lack of statistical adjustment for important covariates [8].

Because of a weak association between self-reported and objectively measured function in patients with chronic pain [9], it is recommended to use both self-reported and objectively measured data for a comprehensive assessment of (work related) illness status [10]. Functional capacity evaluation (FCE) consists of batteries of standardized tests to evaluate an injured worker's functional capacity and ability to perform work-related activities [11]. When FCE results indicate that a worker's functional capacity is less than the job's physical demands, a rehabilitation program can be proposed to improve the ability to return to work [12,13]. FCEs are also used to guide case closure [14,15]. However, the prognostic ability of FCE for (non)-return to work is not known for patients with WAD. This study aimed to: 1) determine the predictive ability of FCE tests to determine future work capacity and 2) to develop a predictive model for work capacity in a cohort of patients with WAD. Our hypotheses were that FCE tests independently predict work capacity in the short term, and that the predictive ability of FCE tests decreases over time.

METHODS

Study design

A prospective cohort study.

Context, subjects, and data collection

Participants were referred from the German-speaking part of Switzerland. They all were insured by the Swiss Accident Insurance Fund (SUVA). SUVA is the largest state owned

accident insurance in Switzerland, which covers occupational and non-occupational injuries for employed individuals, mainly in labor industries, and unemployed job-seeking persons [16]. Injured persons receive compensation up to a maximum of 80% of the previous salary, and medical and vocational assistance. If health status is stabilized but disabilities remain, long term invalidity pensions are refunded by SUVA and the invalidity insurance to the injured person.

Eligible participants were referred for an interdisciplinary rehabilitation assessment at the rehabilitation clinic in Bellikon (Switzerland) by insurance physicians or case managers of SUVA between January 2011 and January 2012. The main reasons for referral included: 1) not regaining full work capacity (WC) within 6–12 weeks after a whiplash injury; 2) exceeding expected healing times; 3) or having plateaued with the provided medical and rehabilitative care. Inclusion criteria for this study were: 1) neck pain due to WAD according the Québec Task Force-Classification (QTF), grade I (pain, stiffness or tenderness without physical signs) or grade II (pain, stiffness, or tenderness with decreased range of motion and point tenderness); 2) WC< 100% of the previous job at the time of the FCE; 3) sufficient German language skills to communicate with the FCE assessors and to respond to questionnaires in German, Albanian, Serbo-Croatian, Italian, Portuguese or Spanish; 4) age 18–65 years; 5) willingness to participate. In total 427 subjects were referred to the interdisciplinary assessment. Of those, 160 were not eligible: 79 (48%) did not have WAD; 46 (28%) had WC of 100%; 17 (10%) had insufficient German language skills or unable to fill out the questionnaires; 6 (4%) had other medical reasons; 5 (3%) had acute comorbidity which limited testing (fracture or severe psychiatric disorder); 3 (2%) were younger than 18 years or older than 65 years; 2 (1%) had WAD grade III–IV; and 2 (1%) were pregnant.

All participants agreed to participate in this study. Ethical approval for this study was granted by the Medical Ethics Committee of Canton Aargau (EK AG 2010/055).

Procedure

A review of the medical history and a physical examination was performed by a rehabilitation physician (approximately 60 min), followed by FCE tests administered by a physiotherapist. After determination of eligibility, patients completed questionnaires and carried out FCE tests (60 min). This was followed by a brief educational intervention and a trial therapy that included a combination of strength exercises, (ergonomic) education and home exercises. The interdisciplinary rehabilitation assessment ended with a face-to-face discussion with the patient about strategies to facilitate recovery. Fitness-for-work certificates or work capacity settlement were explicitly *not* part of this interdisciplinary assessment.

FCE assessors

A sample of 21 physiotherapists (11 female) from the rehabilitation clinic served as FCE assessors. Nineteen had attended a 2-day FCE training course of the Swiss Association of Rehabilitation [17]. Prior to the study all had performed at least ten 1-day FCEs in the previous year (median 30, interquartile range (IQR): 20 to 33) and had a minimum of 1-year experience in work rehabilitation (median 3, IQR: 2 to 3), and a minimum professional practice experience of 1 year (median 5 years, IQR: 3 to 12.5).

Measures

Outcome variable

WC was used as a measure of ability of work. WC was assessed at baseline and 1, 3, 6 and 12 months follow-up. WC was determined by the treating physician, usually a general practitioner, and represents the proportion workability of pre-injury work. Estimation of WC may be determined by suggested measures of WC and based on current national guidelines [18,19]. WC is expressed in a percentage (0–100%). The WC is translated in days or hours modified work. For example, if a worker is deemed WC=50%, he will work for 2.5 days/week or 5 half days/week modified work. The remaining 50% is financially compensated. The WC was obtained from the accident insurance's administrative data. The reliability and validity of the WC determination by physicians is unknown.

Predictor variables

Patients characteristics and probable predictors influencing recovery were recorded prior to FCE: age, gender, body mass index, marital status, mother language, duration since injury, number of previous injury claims, litigation, percentage at work, job contract, education status, and physical work demands. Potential predictor variables were selected based on previous studies and clinical experience [2,4].

The FCE applied in this study (WAD FCE) consisted of 8 tests, based on the Isernhagen Work System (now WorkWell FCE) [11]: handgrip strength right handed, lifting floor to waist, lifting waist to overhead, short two-handed carry, long carry right-handed, overhead working, repetitive reaching right-handed, gait velocity (50 m walking test). Test details are described in the Appendix. Reliability of WAD FCE tests is good to excellent, the tests are safe [20].

Pain intensity was measured with an 11-point numeric rating scale (NRS) ranging from no pain (0) to worst pain (10) [21]. The patient was asked to rate his momentary pain ("pain now"), his worst and his mildest pain during the last week ("pain maximum", "pain minimum"). The NRS has demonstrated reliability and validity in patients with neck pain [22].

Perceived recovery (recovery question, RQ) is a categorical global self-assessment using the question "How well, do you feel, you are recovering from your injuries?", with response options: (1) "all better (cured)," (2) "feeling quite a bit of improvement," (3) "feeling some improvement," (4) "feeling no improvement," (5) "getting a little worse," and (6) "getting much worse." We defined participants as "(somehow) improved" when they reported feeling "all better (cured)" or "feeling quite a bit of improvement", and "feeling some improvement" [23]. The RQ was asked by the rehabilitation physician prior the FCE tests. The RQ was found reliable in patients with WAD [24].

Neck pain-related disability was measured with the Neck Disability Index (NDI). The NDI contains 10 items: pain intensity, personal care, lifting, reading, headaches, concentration, work, driving, sleeping, and recreation. The scale of each item ranges from no disability (0) to total disability (5). Higher NDI scores indicate more disability. The NDI is reliable and deemed valid [25].

The Hospital Anxiety and Depression Scale (HADS) was used to assess the symptom severity of anxiety disorders and depression in non-psychiatric populations. The HADS consists of two scales, one for anxiety and one for depression (A and D scale). Each scale contains seven items, with each item rated from 0 (best) to 3 (worst). The scale scores are calculated by summing the responses up to a maximum score of 21 points (severe case) per scale. Good reliability, validity and excellent screening properties have been reported for the use of the HADS in the general and clinical populations [26].

The Spinal Function Sort (SFS) was used to capture perceived functional ability for work tasks. This questionnaire contains 50 drawings with simple descriptions. Participants rated functional ability for each activity from "unable" (0) to "able" (4). The SFS yields a single rating ranging from 0 to 200, with higher scores indicating better abilities. The scores can be categorized according the work demands as defined by the Dictionary of Occupational Titles (DOT) [27], allowing a comparison between self-reported functional ability and work demands. The SFS has a good reliability and high predictive validity for non-return to work in patients with back pain [28,29].

Submaximal effort determination (S_{ED}) was assessed when a patient stopped a FCE test *before* the FCE rater observed sufficient criteria indicative of maximal weight, or significant functional problems/limitation. The rating of S_{ED} has shown high inter- and intrarater reliability in patients with chronic musculoskeletal pain [30]. A S_{ED} score is the number of FCE items of the total FCE items performed with submaximal effort. A submaximal effort index (SMI) was derived by dividing the total the number of FCE items performed submaximally by the 8 FCE tests performed x 100% (SMI= (n tests submaximal/8)x100%).

Data analysis

Descriptive statistics were computed for baseline patient characteristics and outcome variables. PP- or QQ-plots were used, where appropriate, to test for normality. Bivariate correlations were calculated between FCE tests and WC at follow-up. Linear mixed model were used to determine the predictive value of FCE tests for WC while controlling for confounders.

The analysis included the following steps:

Step 1. All 8 FCE tests and the SMI were entered as predictors in the model with WC at 1, 3, 6, 12 as outcome variables (results not shown; available on request). Regression coefficients with p-value ≥0.1 were *not* considered in following steps of the analysis. Fixed and random effects models were analyzed.

Step 2. In addition of the remaining FCE tests ($p<0.1$) in the model, a random effect coefficient was added to the model which accounted for the effect of predictors which may differ between the participants. We observed an increase of WC over time. Time after baseline assessment was transformed as follows. We took the natural logarithm of the weeks after baseline + 1 week (Ln weeks+1) and that was entered as a predictor in the model.

Step 3. All other potential predictor variables were entered in the model one by one. If the regression coefficients of the remaining FCE tests variables changed by >10%, it was retained for the next step.

Step 4. The remaining FCE tests and the predictor variables were entered in the model simultaneously. The variables were then excluded manually in a backwards selection procedure. Predictors were removed from the model, if the model fit (-2LogLikelihood) did not decrease significantly or the regression coefficient was not significant ($p>0.05$). Interactions between predictors were explored if main effects were significant ($p\leq0.05$). Residuals of the included variables were plotted in a graph to check normality. Data were analyzed in SPSS version 21.0.

RESULTS

Descriptives of the study population

A total of n=267 patients were included. Patient characteristics are displayed in Table 6.1.

Mean WC was at 20.8 (SD 27.6) at baseline and 32.3 (SD 38.4), 51.3 (SD 42.8), 65.6 (SD 42.2), 83.2 (SD 35.0) at 1, 3, 6, 12 months follow-up respectively (Figure 6.1). In a post-hoc analysis we compared the patients WC and corrected for the region of the insurance they were referred to. No regional differences were observed.

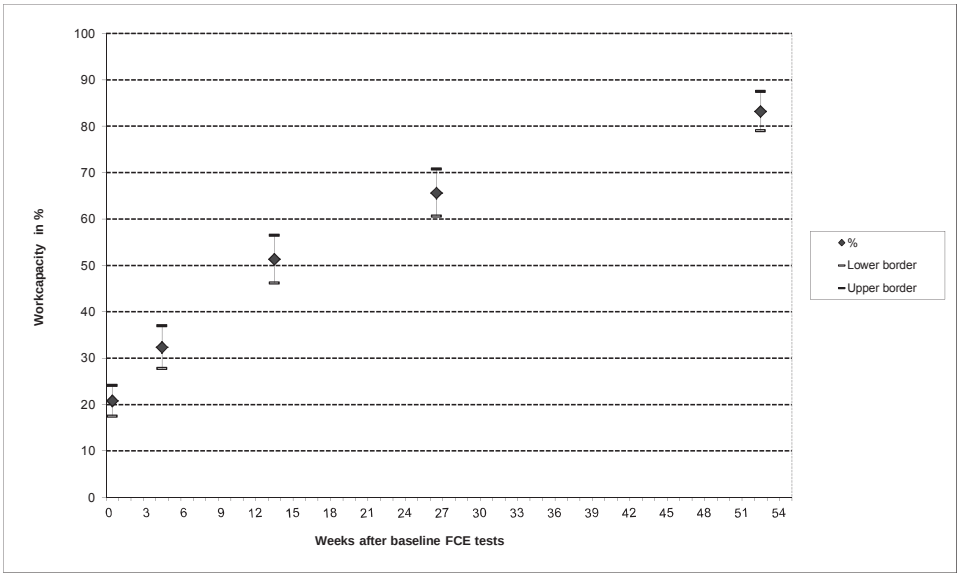


Figure 6.1 Mean workcapacity of the participants at 0, 1, 3, 6, 12 months follow-up.

Table 6.1 Baseline characteristics of the patients (n=267)

Characteristics, unit or scale	
Age (years)*	36.0 (27.0; 44.0)
Gender female, n (%)	106 (39.7)
BMI † *, [2]	26.0 (23.0; 30.0)
Marital status, n (%)	
Married or co-habitation	137 (51.3)
Single	93 (34.8)
Divorced or living separated	36 (13.5)
Other (e.g. widowed)	1 (0.4)
Mother language, n (%)	
(Swiss-)German	131 (49.1)
Other ‡	136 (50.9)
Duration since WAD injury claim opening (days) *	90.0 (71; 122.0)
Number of injury claim openings previous to the current WAD injury *	2.0 (0.0; 5.0)
Attorney involved, n (%)	75 (28.1)
Work capacity in % of the actual or previous work (if jobless)	0.0 (0.0; 50.0)
Work status: job contract §	210 (78.7)
Education ‖, n (%)	
Low	129 (48.3)
Intermediate	132 (49.4)
High	6 (2.2)

Table 6.1 continues on next page

Table 6.1 *Continued*

Characteristics, unit or scale	
Physical work demands ¶, n (%)	
Sedentary to light (<5–10 kg)	89 (33.3)
Medium (11–25 kg)	97 (36.3)
Heavy to very heavy (26 to >45 kg)	81 (30.3)
FCE tests: Mean (SD)	
Hand grip strength (kgF)	33.3 (14.9)
Lifting floor to waist (kg)	18.6 (10.0)
Lifting waist to overhead (kg)	11.2 (5.8)
Short carry two handed (kg)	23.0 (12.1)
Long carry one handed (kg)	16.5 (7.3)
Overhead working (sec) *	166.0 (94; 300)
Repetitive reaching (sec), [1]	82.0 (26.6)
50 m walking test (km/h)	5.1(1.2)
Submaximal effort score (S_{ED} 0–8), number of items *, [1]	2 (0-8)
Self reported measures:	
Pain now (NRS, 0–10) *	5.0 (3.0; 6.0)
Perceived recovery (RQ), n of "somehow improved" ¶¶ (%)	186 (69.7)
Perceived functional ability (SFS, 0–200) *, [5]	136.0 (99.5; 163.0)
Disability (NDI, 0–50), Mean (SD)	23.4 (7.9)
Anxiety (HADS A, 0–21) *	9.0 (6.0; 13.0)
Depression (HADS D, 0–21) *	7.0 (4.0; 10.0)

*if data have a skewed distribution median and an interquartile range (IQR), else mean and SD are provided; [n] = missing data; †BMI = Body mass index; ‡other = 75 (28.1%) Albanian, 23 (8.6%) Serbo-Croatian, 14 (5.2%) Italian, 8 (3.0%) Turkish, 7 (2.6%) Arabic, 3 (1.1%) Portuguese, 1 (0.4%) Spanish, 5 (1.9%) Various; mother language was used as term as a proxy for cultural background or nationality34; § job contract = has a running job contract (≠ jobless) i.e. || Level of education: low = no vocational education, high = vocational education, bachelor or higher education; ¶ Maximum physical work load of material handling tasks in the previous job according to the Dictionary of Occupational Titles (DOT). DOT-categories were merged into three categories; ¶¶ "somehow" was assumed when the patient scored 1–3 on the 6 point scale of the recovery question. SD = standard deviation.

Bivariate analysis

Correlation coefficients between FCE tests and WC decreased over time for most variables (Figure 6.2). The correlation coefficients ranged from r=0.06 (lifting low at 12 months follow-up) to r=0.39 (walking speed at 3 months). Walking speed and S_{ED} showed the highest correlations with WC at follow-up.

Mixed model analysis

The results of the mixed model analysis are presented in Table 6.2.

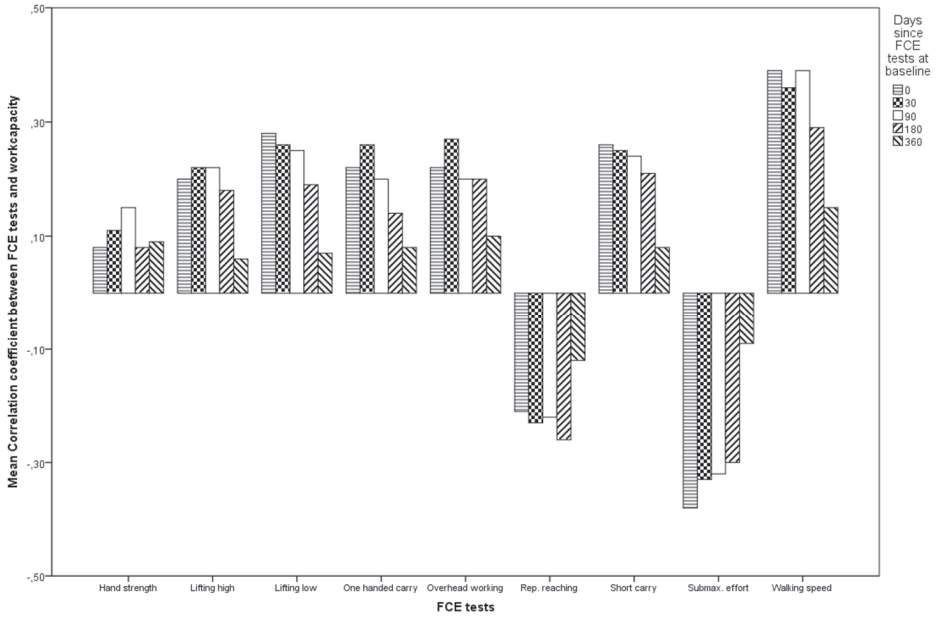


Figure 6.2 Bivariate correlations (Pearson correlations, except for overhead working and submaximal effort Spearman's rank correlation was used) between FCE tests and WC at baseline, 30, 90, 180, 360 days follow-up. For repetitive reaching and submaximal effort score correlations coefficients were negative (negative values were positively transformed in this figure).

The following 3 FCE tests predicted WC: repetitive reaching, gait velocity and the S_{ED} score. The regression coefficients of the three FCE tests in the model decreased from step 2 to step 3 by -0.03 (33%) for repetitive reaching, -6.37 (99%) for walking speed, and -1.66 (91%) for S_{ED} score. In step 4, none of the FCE tests remained significant predictors of future WC. Therefore, FCE tests were excluded from the final model. The final prognostic model included; Ln (weeks+1) (β=23.74), mother language (β=5.49), work capacity at baseline (β=1.01) and self-reported disability (β= -0.20). Time course mediates workcapacity and self-reported disability, as those two interaction terms remained significant. Overall, time course and mother language were the predictors with the highest regression coefficients. To facilitate interpretation of the results of the linear mixed model analysis two clinical examples were calculated (Text Box 6A).

DISCUSSION

We conducted a prospective cohort study to determine the prognostic ability of FCE tests to predict WC and developed a predictive model in a cohort of patients with WAD. Correlation

Table 6.2 Results of the linear mixed model analysis with work capacity in % at 1, 3, 6 and 12 months after baseline assessment as the dependent variable (models steps /evolvement displayed)

Predictors included in step 2 to 4	Coefficients (β)	Standard error (β)	95% CI	P-value
2. Model including effect of time and random effect				
Constant	-5.78	14.39	-34.10 to 22.55	.688
Repetitive reaching	-0.09	0.069	-0.22 to 0.05	.207
Walking speed	6.43	2.01	2.48 to 10.38	.002
Submaximal effort determination (S_{ED}) score	-1.82	0.86	-3.52 to -0.12	.036
Ln (weeks+1)	15.57	0.54	14.51 to 16.63	.000
3. Model including all predictors				
Constant	17.33	15.31	-12.77 to 47.43	.258
Repetitive reaching	-0.04	0.05	-0.15 to 0.06	.428
Walking speed	0.98	1.71	-2.37 to 4.33	.565
Submaximal effort determination (S_{ED}) score	-0.06	0.77	-1.57 to 1.44	.936
Ln (weeks+1)	14.68	0.66	13.39 to 15.97	.000
Work capacity at baseline	0.57	0.05	0.46 to 0.67	.000
Mother language(Swiss-German 1, other 0)	3.48	3.05	-2.51 to 9.48	.254
Number of prior injuries	-0.20	0.32	-0.83 to 0.43	.533
Pain now (NRS)	-0.50	0.74	-1.96 to 0.96	.499
Perceived recovery (RQ)	0.92	2.99	-4.96 to 6.80	759
Perceived functional ability (SFS)	-0.00	0.05	-0.09 to 0.09	.935
Disability (NDI)	-0.41	0.28	-0.96 to 0.13	.132
Anxiety (HADS A)	0.05	0.43	-0.80 to 0.89	.913
Depression (HADS D)	-0.20	0.46	-1.10 to 0.71	.671
4. Model including interaction terms				
Constant	-0.60	7.08	-14.50 to 13.30	.933
Ln (weeks+1)	23.74	2.39	19.04 to 28.44	.000
Work capacity at baseline	1.01	0.07	0.86 to 1.15	.000
Mother language	5.49	2.47	0.64 to 10.34	.027
Disability (NDI)	-0.20	0.26	-0.70 to 0.30	.433
Ln (weeks+1) * Work capacity at baseline	-0.19	0.03	-0.24 to -0.14	.000
Ln (weeks+1) * Disability (NDI)	-0.21	0.09	-0.38 to -0.04	.015

Coefficients (β): unstandardized regression coefficient.

coefficients between FCE tests and WC were <0.4 at baseline, and decreased over the follow-up period. In the multivariate model outcomes of FCE tests do not predict future WC. Our final model suggested that the strongest predictors were time course, mother language, baseline WC, and self-reported disability.

It is recommended to monitor variables with the best predictive capacity in those patients who fail to improve in the transition from acute to chronic [31]. Values of the prognostic variables identified in this study can easily be recorded.

Text Box 6A Two clinical examples are shown in text box for the interpretation of the results

Clinical examples

Formula derived from model 3 in Table 6.2:

Work capacity (%)= 0.60 + (23.74 x Ln (weeks+1) + (1.01 x work capacity base line) + (5.49 x mother language) + (-0.20 x self-reported disability, NDI) + (-0.19 x Ln (weeks+1) x work capacity in % at baseline) + (-0.21 x Ln (weeks+1) x self-reported disability, NDI)

Example A: moderately disabled patient at baseline

Prediction of WC after: 2 weeks from baseline
Work capacity: 60% at baseline
Mother language 1, (Swiss-) German
NDI score: 15

WC = -0.60 + (23.74 x Ln (weeks+1) + (1.01 x work capacity) + (5.49 x mother language) + (-0.20 x self-reported disability NDI) + (-0. 19 x Ln (weeks+1) x work capacity in % at baseline) + (-0.21 x Ln (weeks+1) x self reported disability)

WC = -0.60 + (23.74 x Ln 3) + (1.01 x 60) + (5.49 x 1) + (-0.20 x 15) + (-0.19 x Ln 3 x 60) + (-0.21 x Ln 3 x 15) = **72.2 %**

Example B: severely disabled patient at baseline

Prediction of WC after: 10 weeks from baseline
Work capacity: 10% at baseline
Mother language: 0, non (Swiss-) German
NDI score: 40

WC = -0.60 + (23.74 x Ln (time+ 1) + (1.01 x work capacity) + (5.49 x mother language) + (-0.20 x self-reported disability NDI) + (-0. 19 x Ln (weeks+1) x work capacity in % at baseline) + (-0.21 x Ln (weeks+1) x self reported disability)

WC = -0.60 + (23.74 x Ln 11) + (1.01 x 10) + (5.49 x 0) + (-0.20 x 40) + (-0.19 x Ln 11 x 10) + (-0.21 x Ln 11 x 40) = **33.4%**

Besides WC at baseline, NDI scores and mother language were independent predictors. Whereas, the NDI was predictive also in other populations and settings, the importance of mother langue may be specific for this rehabilitation setting [28,32]. In 2012 populations groups with non-Swiss mother language were employed on average in more physically heavy jobs than their Swiss counterparts [33]. Hence, successful RTW in these populations may be more dependent on physical health and might be more affected by economic fluctuations [34]. Additionally, low health literacy is related to being a non-Swiss mother language [35]. Low health literacy may cause substantial burden to society and the injured person [36]. Understanding the role of language in the development of chronic WAD may be crucial for developing effective work disability prevention programs for patients with WAD.

Predicting return to work in patients with chronic pain is difficult. Lifting tests explain 10-20% of the variance in RTW in patients with musculoskeletal disorders [37]. Some authors reported an explained variance up to 27% [38], while others suggested that, adding FCE tests to self-reported data, would increase the explained variance from 9 to 16% [39]. However, others reported a 10% explained variance questioning the predictive value of FCE tests for RTW in patients with chronic musculoskeletal pain [40,41]. These differences may be explained in differences in the study design, sample size, confounders included, follow-up times, statistical models and corrections, the definition of RTW and social security systems where the studies were performed [8]. This study shows that strength of the correlation between WC and FCE tests is related to the time point after the whiplash injury.

The majority of the patients in this study reached full WC within 12 months follow-up. This is in contrast to others showing that a substantial proportion of patients with WAD (40–60%) still suffer from varying levels of pain and self-reported disability after one year [1,2]. we hypothesized that WC over 12 months may not be indicative of perceived disability. In a posthoc analysis we evaluated the correlation between WC and the available NDI scores at 3 and 12 months (50% of the study sample). The correlations were low ($r<0.3$; WC accounts for 9% of the explained variance of NDI), indicating that disability and WC are related but distinct constructs.

While it may be methodologically correct to study FCE tests separately, in clinical work FCE tests are used in conjunction with medical records, patient interview, musculoskeletal evaluation and job-specific observations [11]. One may argue that predictive value would be higher if RTW can be predicted based on the full clinical package, including FCE tests. Results of this strategy are indeed positive [42,43]. However, methodological challenges accompany this as well [44,45]. Whether a FCE-related interview alone may be an option to FCE tests to predict future WC in patients with WAD, is unknown [46].

Strengths of this study are: the range of known predictive variables consisting self-reported measures, functional capacity tests, and insurance data, a complete dataset of the outcome variable with five measurements over a time period of 12 months [32,47]. Within the analytical approach we controlled for confounders and interactions. The participants, patients and assessors of WC, were blinded to the study hypotheses [8]. Limitations are: The results of the FCE tests were accessible for the treating GP, case manager, physiotherapist and occupational physician, and may have influenced their rating. Co-interventions during the time between 6 to 52 weeks were not controlled for, nor was the type of work, which may be an important confounder for RTW and WC. The accuracy of self-reported measures for disability within a worker's compensation environment can be unreliable [48,49]. However, the alternative (WC) also has shortcomings; it's psychometric properties are unknown and WC is often relying on the patients report and physicians interpretations [50]. Finally, replication studies are needed because the results differ in other populations, contexts and FCE procedures.

CONCLUSIONS

FCE tests performed within 3 months after WAD injury are associated to WC at baseline, but do not predict future WC, whereas time course, mother language, WC at baseline, and self-reported disability do predict future WC. Additionally, interaction between time course, WC at baseline and self-reported disability respectively mediated future WC.

REFERENCES

1. Kamper SJ, Rebbeck TJ, Maher CG, McAuley JH, Sterling M. Course and prognostic factors of whiplash: a systematic review and meta-analysis. Pain 2008;138(3):617-29.
2. Carroll LJ, Hogg-Johnson S, Cote P, van der Velde G, Holm LW, Carragee EJ et al. Course and prognostic factors for neck pain in workers: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine (Phila Pa 1976) 2008;33(4 Suppl):S93-100.
3. Chappuis G, Soltermann B. Number and cost of claims linked to minor cervical trauma in Europe: results from the comparative study by CEA, AREDOC and CEREDOC. Eur Spine J 2008;17(10):1350-7.
4. Scholten-Peeters GG, Verhagen AP, Bekkering GE, van der Windt DA, Barnsley L, Oostendorp RA et al. Prognostic factors of whiplash-associated disorders: a systematic review of prospective cohort studies. Pain 2003;104(1-2):303-22.
5. Carroll LJ, Holm LW, Hogg-Johnson S, Cote P, Cassidy JD, Haldeman S et al. Course and prognostic factors for neck pain in whiplash-associated disorders (WAD): results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine (Phila Pa 1976) 2008;33(4 Suppl):S83-92.
6. Carroll LJ, Ferrari R, Cassidy JD, Cote P. Coping and Recovery in Whiplash-associated Disorders: Early use of Passive Coping Strategies is Associated With Slower Recovery of Neck Pain and Pain-related Disability. Clin J Pain 2014;30(1):1-8.
7. Holm LW, Carroll LJ, Cassidy JD, Skillgate E, Ahlbom A. Expectations for recovery important in the prognosis of whiplash injuries. PLoS Med 2008;5(5):e105.
8. Kamper SJ, Hancock MJ, Maher CG. Optimal designs for prediction studies of whiplash. Spine (Phila Pa 1976) 2011;36(25 Suppl):S268-74.
9. Lin CW, McAuley JH, Macedo L, Barnett DC, Smeets RJ, Verbunt JA. Relationship between physical activity and disability in low back pain: a systematic review and meta-analysis. Pain 2011;152(3):607-13.
10. Annet L, Ilona Z. Review on the validity and reliability of self-reported work-related illness. London; 2012.
11. Isernhagen SJ. Functional capacity evaluation: rational, procedure, utility of the kinesiophysical approach. J Occup Rehabil 1992;2(3):157-68.

12. Henchoz Y, de Goumoens P, Norberg M, Paillex R, So AK. Role of physical exercise in low back pain rehabilitation: a randomized controlled trial of a three-month exercise program in patients who have completed multidisciplinary rehabilitation. Spine (Phila Pa 1976) 2010;35(12):1192-9.

13. Kool JP, Oesch PR, Bachmann S, Knuesel O, Dierkes JG, Russo M et al. Increasing days at work using function-centered rehabilitation in nonacute nonspecific low back pain: a randomized controlled trial. Arch Phys Med Rehabil 2005;86(5):857-64.

14. Oesch PR, Kool JP, Bachmann S, Devereux J. The influence of a Functional Capacity Evaluation on fitness for work certificates in patients with non-specific chronic low back pain. Work 2006;26(3):259-71.

15. Wind H, Gouttebarge V, Kuijer PP, Sluiter JK, Frings-Dresen MH. Effect of Functional Capacity Evaluation information on the judgment of physicians about physical work ability in the context of disability claims. Int Arch Occup Environ Health 2009;82(9):1087-96.

16. Suva. Suva: an overview [Swiss Accident Insurance Fund] 2013. Available at: http://www.suva.ch/english/startseite-en-suva/suva-en-suva/ueberblick-en-suva.htm. Accessed 17.09.2013.

17. Denier-Bont F, Fischer V, Oesch P, Oliveri M. [Functional Capacity Evaluation: Course manual] Bellikon: Verein IG Ergonomie, Swiss Association of Rehabilitation; 2007.

18. Stöckli H, Ettlin T, Gysi F, Knüsel O, Marelli R, Soltermann B. [Diagnostics and therapeutic approach in the chronic phase of whiplash associated disorders]. Schweiz Med Forum 2005;5:1182-7.

19. Fitforwork-swiss. WOCADO [Workcapacity estimation for doctors] 2013. Available at: http://www.fitforwork-swiss.ch/de/projekte.html. Accessed 03.12.2013.

20. Trippolini MA, Reneman MF, Jansen B, Dijkstra PU, Geertzen JH. Reliability and safety of functional capacity evaluation in patients with whiplash associated disorders. J Occup Rehabil 2013;23(3):381-90.

21. Ferraz MB, Quaresma MR, Aquino LR, Atra E, Tugwell P, Goldsmith CH. Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. J Rheumatol 1990;17(8):1022-4.

22. Pool JJ, Ostelo RW, Hoving JL, Bouter LM, de Vet HC. Minimal clinically important change of the Neck Disability Index and the Numerical Rating Scale for patients with neck pain. Spine (Phila Pa 1976) 2007;32(26):3047-51.

23. Carroll LJ, Holm LW, Ferrari R, Ozegovic D, Cassidy JD. Recovery in whiplash-associated disorders: do you get what you expect? J Rheumatol 2009;36(5):1063-70.

24. Ngo T, Stupar M, Cote P, Boyle E, Shearer H. A study of the test-retest reliability of the self-perceived general recovery and self-perceived change in neck pain questions in patients with recent whiplash-associated disorders. Eur Spine J 2010;19(6):957-62.

25. Swanenburg J, Humphreys K, Langenfeld A, Brunner F, Wirth B. Validity and reliability of a German version of the Neck Disability Index (NDI-G). Man Ther 2013;19(1):52-8.

26. Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale. An updated literature review. J Psychosom Res 2002;52(2):69-77.

27. U.S. Department of Labor. The Revised Handbook for Analyzing Jobs. 4th ed. Indianapolis: JIST Works, inc.; 1991.

28. Oesch PR, Hilfiker R, Kool JP, Bachmann S, Hagen KB. Perceived functional ability assessed with the spinal function sort: is it valid for European rehabilitation settings in patients with non-specific non-acute low back pain? Eur Spine J 2010;19(9):1527-33.

29. Borloz S, Trippolini MA, Ballabeni P, Luthi F, Deriaz O. Cross-cultural adaptation, reliability, internal consistency and validation of the Spinal Function Sort (SFS) for French- and German-speaking patients with back complaints. J Occup Rehabil 2012;22(3):387-93.

30. Trippolini MA, Dijkstra PU, Jansen B, Oesch P, Geertzen JH, Reneman MF. Reliability of Clinician Rated Physical Effort Determination During Functional Capacity Evaluation in Patients with Chronic Musculoskeletal Pain. J Occup Rehabil ahead of print DOI 101007/s10926-013-9470-9 2013.

31. Sterling M. Does knowledge of predictors of recovery and nonrecovery assist outcomes after whiplash injury? Spine (Phila Pa 1976) 2011;36(25 Suppl):S257-62.

32. Nordin M, Carragee EJ, Hogg-Johnson S, Weiner SS, Hurwitz EL, Peloso PM et al. Assessment of neck pain and its associated disorders: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine (Phila Pa 1976) 2008;33(4 Suppl):S101-22.

33. Employment and working time. Detailed results form the Swiss Labour Force Survey. 2014. Available at: http://www.bfs.admin.ch/bfs/portal/de/index/themen/03/02/blank/data/03.html. Accessed 10.02.2014.

34. The Swiss Labour Force Survey. 2012. Available at: http://www.bfs.admin.ch/bfs/portal/en/index/themen/03/22/publ.html. Accessed, 10.02.2014.

35. Franzen J, Mantwill S, Rapold R, Schulz PJ. The relationship between functional health literacy and the use of the health system by diabetics in Switzerland. Eur J Public Health 2013;DOIckt202 [pii]10.1093/eurpub/ckt202. 2013.

36. Eichler K, Wieser S, Brugger U. The costs of limited health literacy: a systematic review. Int J Public Health 2009;54(5):313-24.

37. Kuijer PP, Gouttebarge V, Brouwer S, Reneman MF, Frings-Dresen MH. Are performance-based measures predictive of work participation in patients with musculoskeletal disorders? A systematic review. Int Arch Occup Environ Health 2011;85(2):109-23.

38. Branton EN, Arnold KM, Appelt SR, Hodges MM, Battie MC, Gross DP. A short-form functional capacity evaluation predicts time to recovery but not sustained return-to-work. J Occup Rehabil 2010;20(3):387-93.

39. Kuijer PP, Gouttebarge V, Wind H, van Duivenbooden C, Sluiter JK, Frings-Dresen MH. Prognostic value of self-reported work ability and performance-based lifting tests for sustainable return to work among construction workers. Scand J Work Environ Health 2012;38(6):600-3.

40. Streibelt M, Blume C, Thren K, Reneman MF, Mueller-Fahrnow W. Value of functional capacity evaluation information in a clinical setting for predicting return to work. Arch Phys Med Rehabil 2009;90(3):429-34.

41. Gross DP, Battie MC. Functional capacity evaluation performance does not predict sustained return to work in claimants with chronic back pain. J Occup Rehabil 2005;15(3):285-94.

42. Lechner DE, Page JJ, Sheffield G. Predictive validity of a functional capacity evaluation: the physical work performance evaluation. Work 2008;31(1):21-5.

43. Cheng AS, Cheng SW. The predictive validity of job-specific functional capacity evaluation on the employment status of patients with nonspecific low back pain. J Occup Environ Med 2010;52(7):719-24.

44. Reneman MF, Dijkstra PU. Predictive validity of FCE? Work 2009;32(1):105-6; author reply 7-8.

45. Reneman MF, Soer R. Was predictive validity of a job-specific FCE established? J Occup Environ Med 2010;52(12):1145; author reply -6.

46. Gross DP, Asante AK, Miciak M, Battie MC, Carroll LJ, Sun A et al. A Cluster Randomized Clinical Trial Comparing Functional Capacity Evaluation and Functional Interviewing as Components of Occupational Rehabilitation Programs. J Occup Rehabil. DOI 10.1007/s10926-013-9491-4. 2014.

47. Sterling M, Carroll LJ, Kasch H, Kamper SJ, Stemper B. Prognosis after whiplash injury: where to from here? Discussion paper 4. Spine (Phila Pa 1976) 2011;36(25 Suppl):S330-4.

48. Gauthier N, Sullivan MJ, Adams H, Stanish WD, Thibault P. Investigating risk factors for chronicity: the importance of distinguishing between return-to-work status and self-report measures of disability. J Occup Environ Med 2006;48(3):312-8.

49. Don AS, Carragee EJ. Is the self-reported history accurate in patients with persistent axial pain after a motor vehicle accident? Spine J 2009;9(1):4-12.

50. Rainville J, Pransky G, Indahl A, Mayer EK. The physician as disability advisor for patients with musculoskeletal complaints. Spine (Phila Pa 1976) 2005;30(22):2579-84.

APPENDICES

FCE test procedure and patient instructions

Patients were briefly instructed on how to perform each test. The assessor first gave a single demonstration of each test. Lifting tests were commenced with a light weight. Patients were then asked to perform the tests to their maximum ability. Weights lifted incrementally increased according to a patient's performance, using weights of 2.5 and 5 kilograms. To determine the physical effort level, testers used observational criteria indicating physical demand. Testing could be terminated for four reasons: the patient stopped because of, for example, pain; the assessor deemed testing to have become unsafe based on biomechanical criteria; heart rate exceeded 85% of the age-related maximum (220 minus age of patient); or a predefined time limit was reached. If a patient stopped the lifting waist to overhead test before the criteria for maximum level of demand was observed the highest weight in kilogram that the patient was willing to lift five times was recorded.

FCE test descriptions

Isometric hand grip strength

Isometric hand grip strength was measured in a seated position. The subjects held their shoulder adducted without internal or external rotation, elbow flexed at approximately 90° and the forearm and wrist in neutral position. Grip strength of the right and left hand was measured in a three-trial procedure while maintaining in a hand dynamometer in a one handgrip position (Jamar PC 5030, Preston Corporation, 1994). An average amount of kgF was scored.

Material handling tests

All lifting tests were executed with a wooden crate (40 x 30 x 26 cm) of 2.5 kg, and four to five weight increments of 2.5 kg or 5 kg each were used until the maximum amount of weight was reached. Maximum performance was recorded in kg.

Lifting floor to waist was measured after five lifts of the crate from floor to table and vice versa (time limit <90 s): hands remained on the crate during the test.

Lifting waist to overhead was measured during lifting of the crate from table to crown in standing position, and vice versa.

Two-handed carrying of a crate for a short distance was measured after five carries of 1.5 m distance at waist height. Hands remained on the crate during the test.

The one-handed carrying of a wooden crate for 15 m within 60 sec began with the right hand and thereafter the left hand.

Overhead work test

Overhead working was performed standing with hands at crown height for manipulation of nuts and bolts. The time that the position was held was recorded (sec).

Repetitive reaching test

Repetitive reaching was determined by fast horizontal movements of the upper extremity in a sitting position. Marbles were removed from bowls at arm length distance at table height from left to right and vice versa, with right and then left arm. The time taken to remove 30 marbles was recorded (sec).

50 m walking test

The walking test was executed on a 50 m-distance track. Participants were asked to walk as fast as possible. The instruction was: "Pause is allowed. Do not run!" The time taken to walk for 50 m was measured (sec), and km/h was calculated.

Chapter 7

Measurement properties of the Spinal Function Sort in patients with sub-acute whiplash-associated disorders

Maurizio A. Trippolini
Pieter U. Dijkstra
Jan H. B. Geertzen
Michiel F. Reneman

Journal of Occupational Rehabilitation (Accepted, pending minor revision).

ABSTRACT

Purpose: To extensively analyze the measurement properties the Spinal Function Sort (SFS) in patients with sub-acute whiplash-associated disorders (WAD).

Methods: Three-hundred-two patients with WAD were recruited from an outpatient work rehabilitation center. Internal consistency was assessed by Cronbach's α. Construct validity was tested based on 8 a priori hypotheses. Structural validity was measured with principal component analysis (PCA). Test-retest reliability and agreement was evaluated in a sub sample (n=32) using interclass correlation coefficient (ICC) and limits of agreement (LoA). The predictive validity of SFS for future work status at 1, 3, 6 and 12 months follow-up was determined by area under the curve (AUC) of receiver operating characteristics. Non-return to work (N-RTW) was defined with two cut-off points: workcapacity <50% and <100%.

Results: N-RTW decreased from 50%, 1 month follow-up, to 14%, 12 months follow-up. Cronbach's α was 0.98, PCA revealed evidence for unidimensionality. ICC was 0.86, LoA was ±33 points. Seven out of 8 eight hypotheses for construct validity were not rejected. AUC reduced with a longer follow-up from 0.71 for 1 month to 0.68 at 12 months, for cut-off point <50%. For cut-off point <100% these values were 0.71 and 0.59.

Conclusions: In patients with sub-acute WAD test-retest reliability, internal consistency, construct- and structural validity of the SFS were adequate. LoA were substantial. Sensitivity to accurately predict N-RTW was poor. The predictive validity of the SFS for N-RTW of patients with sub-acute WAD from an outpatient work rehabilitation setting was only sufficient for the short term (1 month).

INTRODUCTION

Self-report questionnaires have been developed for many types of health conditions, some for use in occupational rehabilitation. One of the reasons for their popularity is the relative efficiency of data collection. In limited, a broad array of data can be collected about the functional impairments, limitations, and psychological status experienced by the evaluee. This information can be very useful for planning return to work interventions.

However, disability questionnaires have important limitations for use in European occupational rehabilitation settings. The first is that the use of self-reported measures depends on the literacy and linguistic skills of an evaluee which may be limited in evaluees with different cultural backgrounds i.e. mother languages [1]. The second is that most disability instruments do not have a work-related point of reference, but consider an unlimited spectrum of activities. Whether or not the evaluee can actually lift 15 kg at work, for example, is still unknown after filling in the questionnaire. These limitations may be overcome by using a picture-based questionnaire such as the Spinal Function Sort (SFS) [2]. The SFS is a self-report measure of tasks and activities that includes a picture to each item [3]. The items are linked to demonstrable physical ability. This measure is used in conjunction with a functional capacity evaluation (FCE) to cross-reference self-reported abilities with measured abilities (functional capacity) [4].

In patients with chronic low back pain (CLBP) the SFS has revealed good clinical practicality, reliability and high predictive validity for non-return to work in various settings and countries [5-8]. Although, the SFS is used in occupational health for other health conditions as well, reliability and (predictive) validity of SFS other than CLBP is unknown. Furthermore, it is not reported whether the SFS performs differently in samples with a shorter disease duration.

Hence, the aim of this study was to test measurement properties of the SFS by assessing internal consistency, test-retest reliability, agreement, construct validity and predictive validity for work status of the SFS in patients with sub-acute WAD.

METHODS

Subjects, procedure and context

Subjects

This study was embedded within usual care of an outpatient work rehabilitation setting. From January 2011 to January 2012 eligible participants were referred for an interdisciplinary rehabilitation assessment at the rehabilitation clinic in Bellikon (Switzerland) by insurance

physicians or case managers of Swiss Accident Insurance Fund (SUVA). Participants were from the German-speaking part of Switzerland. The main reasons for referral included: 1) not regaining full work capacity (WC) within 6–12 weeks after a whiplash injury; 2) exceeding expected healing times; 3) or having plateaued with the provided medical and rehabilitative care. Inclusion criteria were: injured workers with WAD related neck pain, Grade I or II according the Québec Task Force Classification with reduced working capacity of their actual job. They were within 6–12 weeks after initial injury, and received worker's compensation benefits.

Ethical approval for this study was granted by the Medical Ethics Committee of the Canton Aargau (EK AG 2010/055). Patients gave consent that their data was used for research purpose.

Procedure

At base line a review of the medical history and a physical examination was performed by a rehabilitation physician (approximately 60 min), followed by FCE tests administered by a physiotherapist. After determination of eligibility, patients completed questionnaires and carried out FCE tests (60 min). Fitness-for-work certificates or work capacity settlement were explicitly *not* part of this interdisciplinary assessment.

Context

All participants were insured by SUVA, the largest state owned accident insurance in Switzerland. SUVA covers costs for occupational and non-occupational injuries for employed individuals and unemployed job-seeking persons [9]. Injured persons receive compensation up to a maximum of 80% of the previous salary, and medical and vocational assistance. Invalidity pensions can also be refunded by SUVA to the injured person.

Measures

SFS

The SFS was used to measure self-reported functional ability to perform work-related tasks and activities of daily life that involve the spine. The SFS contains 50 drawings with simple descriptions (Item example in the Figure 7.1). Patients rated their functional ability for each activity on a 5-point Likert scale: "able" (4), to "restricted" (1, 2, 3) or "unable" (0). The SFS yields a single rating ranging from 0 to 200, with higher scores indicating more or better abilities. The scores can be categorized according the work demands as defined by the Dictionary of Occupational Titles (DOT) [10]. SFS scores have been adapted to the DOT categories previously as follows [5]: SFS score <100 ≈ minimal work demands, 100–124 ≈ sedentary work (<5kg), 125–164 ≈ light work (5–10 kg), 165–179 ≈ medium heavy work (10–25kg),

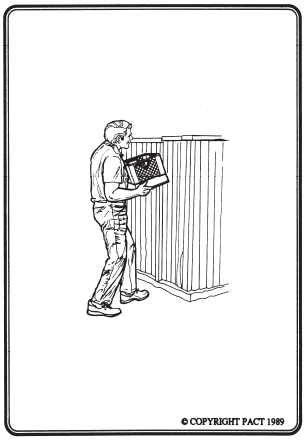


Figure 7.1 Item 14 of the Spinal Function Sort (SFS) questionnaire: Lift a 10 kg milk crate from the floor to eye-level. © Copyright: PACT 1989. All rights reserved.

180–194 ≈ heavy work (25–45), >195 ≈ very heavy work (>45 kg) These categories allow a comparison between the self-reported functional ability and work demand. For test-retest reliability of the SFS a sample of patients was tested twice within a week after baseline.

Work capacity (WC)

To determine the predictive validity for future work status, the WC was used as a measure of ability of work. The WC was obtained from the accident insurance's administrative data. It was determined at 1, 3, 6 and 12 months after baseline by the treating physician, usually a general practitioner, and represents the proportion workability of pre-injury work, expressed in a percentage (0–100%). Estimation of WC was based on proposed WC-estimation forms and recommendations [11,12]. The WC was transformed in days or hours modified work. For example, if a worker is deemed WC=50%, he will work for 2.5 days/week or 5 half days/week modified work. The remaining 50% is financially compensated. The reliability and validity of the WC determination is unknown.

FCE

FCE is a standardized battery of functional tests that intend to measure a patient's safe physical ability for work related activity [13]. For the purpose of this study four lifting tests were analyzed: lifting floor to waist, lifting waist to overhead, short two handed carry, long one-handed carry (right). Patients were asked to perform the test to their maximum ability. The tests have good reliability and acceptable agreement in patients with WAD [14].

Pain

Pain intensity was measured with an 11-point numeric rating scale (NRS) ranging from no pain (0) to worst pain (10) [15]. The patient was asked to rate his momentary pain ("pain now"). The NRS is a commonly used scale with proven reliability and validity in patients with neck pain [16].

Disability

Neck pain-related disability was measured with the Neck Disability Index (NDI) [17]. The NDI contains 10 items: pain intensity, personal care, lifting, reading, headaches, concentration, work, driving, sleeping, and recreation. The scale of each item ranges from no disability (0) to total disability (5). A higher score indicates more severe self-reported disability. The NDI is reliable and valid in several languages and settings [17,18] .

Mental distress

The Hospital Anxiety and Depression Scale (HADS) was used to assess the symptom severity of anxiety disorders and depression in non-psychiatric populations. The HADS consists of two scales, one for anxiety and one for depression (A- and D-scale respectively). Each scale contains 7 items, with each item rated from 0 (best) to 3 (worst). The scale scores are calculated by summing the responses to the items up to a maximum score of 21 points (severe case) per scale. A higher score indicates more severe anxiety or depression. Good reliability, validity have been reported for the use of the HADS in the general and various clinical populations [19,20].

Data analysis

Normal distribution was visually assessed using P-P plots and tested with the Kolmogorov-Smirnov and the Shapiro-Wilk tests. Floor and ceiling effects for the SFS were considered to be present if more than 15% of participants achieved the lowest or highest possible score of the items [21].

Internal consistency

Internal consistency was assessed by item-to total correlations and Cronbach's alpha. Optimal consistency for measurements at group level was considered when alpha value is between 0.7 and 0.9. Values <0.7 may be indicative for items measuring different traits, values >0.9 may be indicative for item redundancy [22].

Unidimensionality

The unidimensionality of the 50 SFS items was measured with principal component analysis (PCA) with Kaiser normalization and Varimax rotation. An Eigenvalue criterion of 1.0 was used for the factor analysis. Unidimensionality was assumed when ratio of the first to the second factor was 3:1 [23].

Test-retest reliability and agreement

Test-retest reliability was expressed as an Interclass Correlation Coefficient (model 1; one-way random) (ICC). ICC was interpreted as follows: ICC≥0.90 is excellent; good when ICC was between 0.75 and 0.90; moderate when ICC was between 0.50 and 0.75; and poor when ICC≤0.50. ICCs were acceptable when ICC≥0.75, and the lower boundary of the 95% confidence interval of the ICC≥0.50 [24]. Agreement was expressed in limits of agreement (LoA) (mean difference ± 1.96 x standard deviation of mean difference) [25].

Construct validation: hypothesis testing

Eight predefined hypothesis on the strength of the association of SFS and four FCE lifting tests, NDI, Pain NRS, and HADS A+D are displayed in Text Box 7A. The strength of the association is expressed in the absolute value of the correlation coefficient. Associations were calculated using Spearman rank correlation coefficient and interpreted as follows: 0.00–0.25 little if any ("not correlated"); 0.26–0.49 low or weak; 0.50–0.69 moderate; 0.70–0.89 high or strong; 0.90–1.00 very strong correlation [26]. The SFS was considered valid, when 7 out of 8 hypotheses (≥80%) of the a priori hypotheses were not rejected [27].

Text Box 7A Eight hypotheses for examining construct validity of the Spinal Function Sort

	Reference test	The validity is not rejected if the strength of the relationship of SFS with	r cut-off values
1	Four lifting tests [a]	functional lifting tests is moderate to high	$0.50 \leq \lvert r \rvert \leq 0.89$
2	Self-reported disability (NDI)	self-reported disability is moderate	$0.50 \leq \lvert r \rvert \leq 0.70$
3	Pain now (NRS)	pain is low or weak	$0.25 < \lvert r \rvert < 0.50$
4	Anxiety (HADS A)	anxiety is low or weak	$0.25 < \lvert r \rvert < 0.50$
5	Depression (HADS D)	depression is low or weak	$0.25 < \lvert r \rvert < 0.50$

[a] Lifting tests include lifting floor to waist (kg), lifting waist to overhead (kg),) short carry two-handed (kg), one-handed carrying right (kg). $\lvert r \rvert$ = Correlation Coefficient, absolute value. The direction of the association depends on the scoring of the reference measure.

Predictive validity for work status at 1, 3, 6 and 12 months

Sensitivity and specificity, positive predictive value as well as likelihood ratio of a positive test were calculated to evaluate the predictive validity of the SFS items at baseline for work capacity at 1, 3, 6 and 12 months after baseline assessment. In a setting of injured workers, who are in a transition phase from acute to chronic disorder, the aim is to identify those patients with a high probability of not returning to work in order to target specific rehabilitation interventions to those patients. Thus our outcome was not return to work and using as reference test WC, applying two cut-off points i.e. WC <50%, or WC <100%. These two cut-off points were determined based on distribution-plots of WC. The index test was the SFS. Sensitivity was defined as the proportion of patients, identified for different DOT categories based on the SFS score, not have returned to work (N-RTW). Specificity was defined as the proportion of patients, identified for different DOT-categories based on the SFS score, who did return to work. The positive predictive value for N-RTW was calculated as the percentage of patients within a DOT category that were correctly identified not to have regained full work capacity. Likelihood ratio was calculated as Sensitivity / 1- Specificity. Based on a previous study, it was expected that "minimal", perceived ability (SFS score <100, less than sedentary work) score would have a high positive predictive value in identifying those patients who would N-RTW at follow-up times [5]. Receiver operating characteristic (ROC) curves were drawn and area under the curve (AUC) was calculated. The AUC has a maximum value of 1.0, indicating a perfect predictive validity which is reached if the curve lies in the upper-left corner; a value of 0.5, represented by the diagonal, means that the measurement instrument cannot distinguish between patients N-RTW or RTW. An AUC of at least 0.70 is considered "appropriate" [28]. As a cut off indicating statistical significance $p<0.05$ was used. All analyses were performed using SPSS (Statistical Package for Social Sciences, Version 21).

RESULTS

Patients

From January 2011 to January 2012, 313 subjects were eligible based on the inclusion criteria. Seven SFS scores were missing. In the construct validity study 306 subjects were included. From this sample 302 were included in the study on the predictive validity of the SFS because 4 patients no follow-data on WC was available (Table 7.1). For the test-retest reliability 32, 11 females, 21 males, mean age 39.6 years, were assessed twice within a week. The patients characteristics of the test-retest study are reported elsewhere [14].

Table 7.1 Characteristics of the patients (n=302)

Characteristics, unit or scale	
Age (years)	36.1 (11.5)
Female, n (%)	130 (43.0)
Marital status, n (%)	
Married or co-habitation	155 (51.3)
Single	104 (34.4)
Divorced or living separated	41 (13.6)
Other (e.g. widowed)	2 (0.7)
Mother language, n (%)	
Swiss (-German)	157 (52.0)
Albanian	79 (26.2)
Serbo-Croatian	23 (7.6)
Italian	16 (5.3)
Turkish	10 (3.3.)
Arabic	7 (2.3)
Portuguese	3 (1.0)
Spanish	1 (0.3)
Other [a]	6 (2.0)
Duration since WAD injury claim opening (days) *	91.0 (72; 125.0)
Attorney involved, n (%)	82 (27.2)
Work status: job contract, n (%)	240 (79.5
Education [b], n (%)	
Low	142 (47.0)
Intermediate	152 (50.3)
High	8 (2.6)
FCE tests	
Lifting floor to waist (kg)	19.4 (10.1)
Lifting waist to overhead (kg)	10.7 (5.8)
Short carry two-handed (kg)	23.7 (12.2)
Long carry one handed (kg)	16.9 (7.6)
Self-reported measures (Scoring range)	
Pain now (NRS, 0–10) *	5.0 (3.0; 6.0)
Perceived functional ability (SFS, 0–200) *	141.0 (103.00; 167.0)
Disability (NDI, 0–50)	22.4 (8.3)
Anxiety (HADS A, 0–21) *	9.0 (5.0; 12.0)
Depression (HADS D, 0–21) *	7.0 (3.0; 10.0)

* If data have a skewed distribution median and an interquartile range, else mean and standard deviation are provided; [a] Other = 1 Polish, 1 Dutch, 1 unknown; [b] Level of education: low = no vocational education, high = vocational education, bachelor or higher education.

Internal consistency, ceiling effects

Internal consistency was Cronbach's alpha 0.98. Removing 50% of the items (even or uneven items), resulted in alpha values of 0.97. Ceiling effects were not present, except in items 45–48. The item to total correlation was <0.20 in item 45–48. These four items displayed very heavy material handling tasks (>45 kg). In a post hoc analysis, Cronbach's alpha values were unchanged when removing item 45–48. All other items showed item to total correlations >0.30.

Unidimensionality

Correlations coefficients between each of the SFS were in the majority >0.3. PCA with fixed factors showed the presence of 6 components with Eigenvalues exceeding 1, explaining 55.3%, 8.2%, 4.6%, 3.2%, 2.3% and 2.1% of the variance, respectively. The inspection of the scree plot revealed 2 components. For the interpretation of the components Varimax rotation was executed. The rotated solution revealed the presence of a mixed structure with two components showing a number of strong loadings. The items 45–48 loaded on a different component. The ratio from the first to the second Eigenvalue was 6.87, indicating reasonable evidence for unidimensionality.

Test-retest reliability and agreement

The test-retest reliability measured with the ICC was 0.86 (95% CI: 0.71; 0.93). For the 32 patients in the reliability study, mean SFS scores for test and retest were 146.4 (mean, SD 32.1), and 146.6 (mean, SD 37.2) respectively. Mean difference in SFS score between test and retest was 0.2 (SD 16.9, p=0.943). Hence LOA were 0.2 ± 33 points Variances were not related to the magnitude of the score. A highly influential patient with a difference of 62 units between tests was detected. LoA calculated without that patient were -23.2 and 27.7 with a mean difference of 2.2 (Figure 7.2).

Construct validity

Construct validation: hypothesis testing

Spearman rank correlations coefficient between the SFS and FCE tests were for lifting floor to waist, lifting waist to overhead; short two-handed horizontal carry, one-handed carry right: 0.68, 0.61, 0.70 and 0.64, respectively. Correlations between the SFS and disability, pain, anxiety and depression were: -0.62, -0.49, -0.49 and -0.52, respectively. All correlations were significant (p-value <0.01). Seven of 8 hypotheses were not rejected. Correlations between

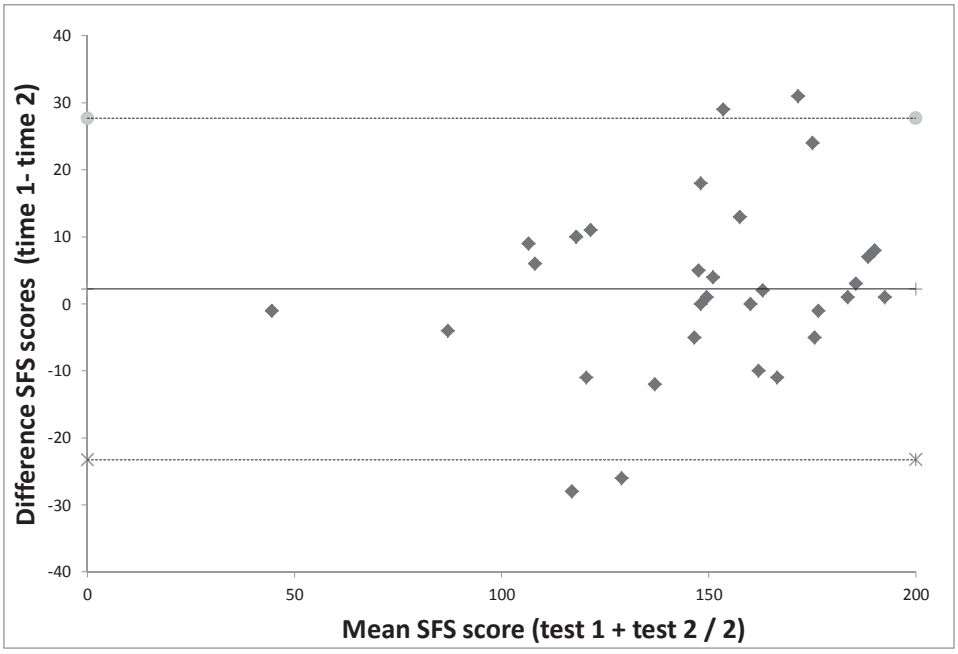


Figure 7.2 Bland-Altman plot of the SFS scores. The middle line represents the mean difference between the two tests. ● represents the upper, and * the lower limit of agreement, i.e. mean difference + 1.96 SD of the differences and mean difference - 1.96 SD of the differences, respectively. An outlier with a difference in SFS scores of 62 is not shown.

SFS and work-related lifting tests was moderate to high (0.61–0.70). Depression showed a slightly stronger correlation than hypothesized.

Predictive validity for work status at 1, 3, 6 and 12 months follow-up

Sensitivity of the SFS scores transformed into DOT categories for N-RTW at 1, 3, 6 and 12 months ranged between and 0.37 and 0.98 when using the cut-off value of <50% WC between 0.28 and 0.98, with the cut-off <100% respectively (Table 7.2). Sensitivity was substantially higher in the DOT-transformed categories "light" to "very heavy" than in the "sedentary" to "minimal" categories (Table 7.2). The likelihood ratio for a positive test for N-RTW at 1, 3, 6 and 12 months decreases from 4.64 to 0.96 for the cut-off value <50% WC, and from 4.32 to 0.79 for the cut-off value of <100% WC.

SFS score can be dichotomized into scores <100 and scores ≥100 points. Patients with scores <100 perceive themselves as having minimal working ability. With this dichotomized scores, Sensitivity for N-RTW with the cut-off of WC<50% ranged over time between 0.37 and 0.41,

Table 7.2 Predictive validity of DOT-transformed SFS categories for non-return to work at 1, 3, 6 and 12 months of follow-up

DOT categories (SFS score adapted)	N-RTW	RTW	Sens	Spec	+PV	Lr+	N-RTW	RTW	Sens	Spec	+PV	Lr+
	WC-Cut-off: 0–49%	WC-Cut-off: 50–100%					WC-Cut-off: 0–99%	WC-Cut-off 100%				
1 month follow-up												
Minimal (0–99)	55	12	0.37	0.92	0.82	4.64	62	5	0.28	0.94	0.93	4.32
Sedentary (100–124)	26	19	0.54	0.80	0.72	2.65	41	4	0.46	0.88	0.92	3.99
Light (125–164)	43	66	0.83	0.36	0.56	1.30	72	37	0.78	0.41	0.79	1.32
Medium (165–179)	13	26	0.91	0.19	0.53	1.13	25	14	0.89	0.23	0.77	1.16
Heavy (180–194)	9	20	0.97	0.06	0.51	1.03	19	10	0.98	0.10	0.76	1.09
Very heavy (195–200)	4	9					5	8				
3 month follow-up												
Minimal (0–99)	43	24	0.41	0.88	0.64	3.41	56	11	0.32	0.91	0.84	3.69
Sedentary (100–124)	17	28	0.58	0.74	0.54	2.20	29	16	0.49	0.79	0.81	2.28
Light (125–164)	33	76	0.89	0.35	0.42	1.38	55	54	0.80	0.36	0.67	1.25
Medium (165–179)	6	33	0.95	0.19	0.39	1.17	17	22	0.90	0.19	0.64	1.11
Heavy (180–194)	3	26	0.98	0.06	0.35	1.04	14	15	0.98	0.07	0.59	1.05
Very heavy (195–200)	2	11					4	9				

Table 7.2 continues on next page

Table 7.2 *Continued*

DOT categories (SFS score adapted)	N-RTW	RTW	Sens	Spec	+PV	Lr+	N-RTW	RTW	Sens	Spec	+PV	Lr+
	WC-Cut-off: 0–49%	WC-Cut-off: 50–100%					WC-Cut-off: 0–99%	WC-Cut-off 100%				
6 month follow-up												
Minimal (0–99)	28	39	0.38	0.83	0.42	2.25	45	22	0.34	0.87	0.67	2.67
Sedentary (100–124)	12	33	0.55	0.69	0.36	1.74	21	24	0.50	0.73	0.59	1.87
Light (125–164)	26	83	0.90	0.32	0.30	1.34	42	67	0.82	0.34	0.49	1.25
Medium (165–179)	4	35	0.96	0.17	0.27	1.16	8	31	0.66	0.16	0.45	0.79
Heavy (180–194)	1	28	0.97	0.05	0.25	1.02	10	19	0.96	0.05	0.44	1.01
Very heavy (195–200)	2	11					5	8				
12 month follow-up												
Minimal (0–99)	15	52	0.37	0.80	0.22	1.84	21	46	0.33	0.81	0.31	1.73
Sedentary (100–124)	6	39	0.51	0.65	0.19	1.47	10	35	0.49	0.66	0.28	1.45
Light (125–164)	13	96	0.83	0.28	0.15	1.16	19	90	0.79	0.28	0.23	1.11
Medium (165–179)	4	35	0.93	0.15	0.15	1.09	5	34	0.87	0.14	0.21	1.02
Heavy (180–194)	0	29	0.93	0.04	0.13	0.96	5	24	0.95	0.04	0.21	0.99
Very heavy (195–200)	3	10					3	10				

N-RTW: not return to work based on the WC; RTW: return to work based on the WC; Spec specificity, Sens sensitivity, +PV positive predictive value; Lr+ likelihood ratio of a positive test; DOT: Dictionary of Occupational Titles; SFS: Spinal Function Sort.

and specificity (=RTW) ranged between 0.80–0.92. For the cut-off of WC<100%: sensitivity for N-RTW ranged over time between 0.28 and 0.34 and specificity (=RTW) ranged between 0.81 and 0.94 (based on data in Table 7.2, separately available on request).

All ROC curves are displayed in Figure 7.3. The AUC reached the cut-off for "acceptable" of >0.70 only at 1 month follow for both WC cut-offs used.

DISCUSSION

The aim of the study was to extensively analyze measurement properties of the SFS in patients with WAD 6–12 weeks after injury. The majority (7 out of 8) of the a-priori defined hypotheses

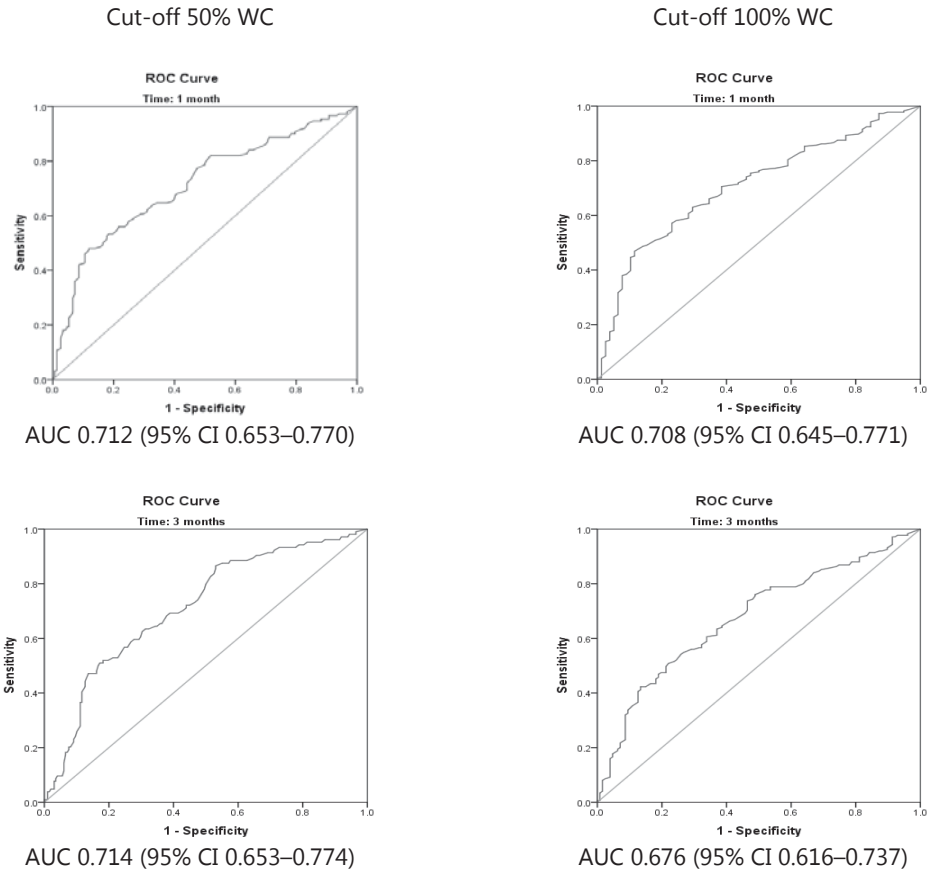


Figure 7.3A ROC curve of SFS total score at baseline with cut-off values of work capacity 50% or 100% at 1 month (first row) and 3 month (second row) follow-up to predict non return to work. WC: workcapacity; AUC: area under the curve; CI: confidence interval.

for construct validity were not rejected. The SFS test structure was confirmed by a distinct factor loading. Test-retest reliability was good, however measure of error (LoA values) on an individual level were large relative to the scale range. Predictive validity of the SFS based on the AUC was acceptable only at 1 month. The SFS scores for the DOT-transformed categories "minimal" to "sedentary" workload were not able to identify those who will N-RTW (low sensitivity). The positive likelihood ratio for N-RTW was sufficient only for the categories "minimal" to "sedentary" for both cut-off WC <50% and WC <100%.

The SFS can, based on the measurement properties evaluated in this study, be recommended for clinical and research applications in patients in an occupational setting with sub-acute WAD and with different cultural backgrounds. Clinicians should be aware of the large measurement error of the SFS when making recommendations on individual level. The scores of the SFS

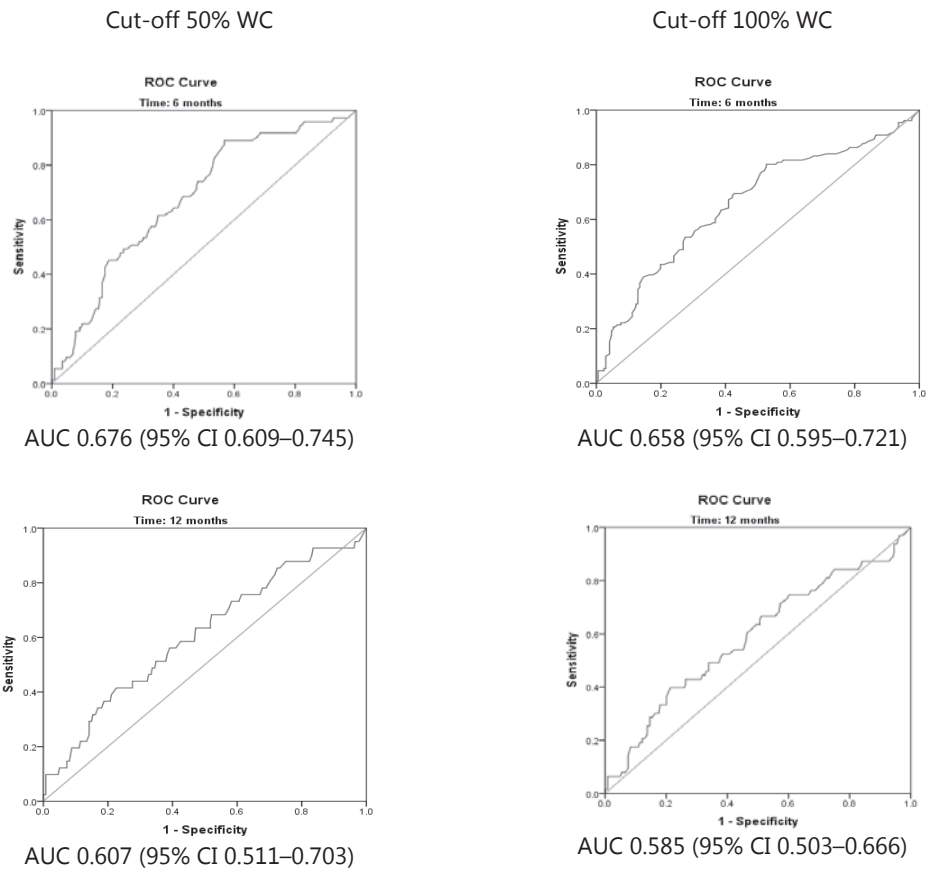


Figure 7.3B ROC curve of SFS total score at baseline with cut-off values of workcapacity 50% or 100% at 6 month (first row) and 12 month (second row) follow-up to predict non return to work.

may assist to predict N-RTW especially for medium, heavy and very heavy DOT categories. Application of the SFS may be a practical alternative or addition to other instruments with sufficient measurement properties. Practicality can be enhanced when half the items are removed. Further research should analyze if even more items can be removed (Cronbach's α of half the SFS items is 0.97, indicating that item redundancy is still apparent).

The SFS scores in our sub-acute sample was substantially higher (mean 133 points, SD 42.7) than in two other validation studies with chronic low back pain patients in Europe (mean 105 points, SD 46.1), and in Australia (mean 116, SD 40.8) [5,8]. A very high Cronbach's α was found, which is in line with previous validation studies [5,6,8]. High internal consistency may be partly determined by a large number of items [29]. These high alpha values are indicative for item redundancy. In a sensitivity analysis we calculated Cronbach's α and PCA values with half of the SFS items, with minimal changes in consistency and dimensionality. From a statistical point of view, half of the SFS items could be omitted, reducing the time requirement to fill out the questionnaire to 5 Min. (now, 10–15 Min.). In agreement with previous studies, four items, with very heavy lifting tasks, could be removed without affecting the measurement properties of the SFS [5,6]. Our results concerning reliability measured with ICC 0.80 are lower than two reliability studies 0.89 and 0.98 respectively [6,8]. The LoA values found in a rehabilitation setting in the French-speaking area of Switzerland were ±11 while in the German-speaking area the values were ±27, whereas our results were ±33 [6]. In the studies of the German speaking sample the SFS was part of case-closure FCE setting to define fitness-for-work, whereas in the French-speaking sample this was not the case [5,6]. One reason for the differences in reliability and agreement may be the difference in interval between test and retest; 2–3 days compared to 7 days in our study. Another reason may be that our patients were in a sub-acute stage of WAD which may change more on a daily basis compared to chronic patients. The ability to predict N-RTW in our study was substantially lower than in a sample of patients with CLBP [5] although follow-up times were similar. Albeit some similarities, the work rehabilitation setting and large proportion of blue collar workers with Non-Swiss cultural background, several other reasons may explain these differences.

First, the proportion of patients who did N-RTW was substantially lower at 3 and 12 month follow-up in our study sample compared in patients with CLBP with rates between 34% to 16%, and 62% and 54% respectively. This may be due the fact that the CLBP patient had on average a significantly longer duration of 200 days off work, compared to 90 days in this study. Therefore, a smaller proportion of WAD patients is expected to N-RTW due to the benign natural course of the disorder despite perceived disability [30]. Further, we used WC data from the physician and the insurance. Moreover, legal regulations in Switzerland recently changed allowing to close claims of patients with WAD within the first 1 or 2 years which is

not the case in CLBP [31]. These changes may have influenced N-RTW rates in patients with WAD which depend on the legal jurisdictions [32]. Hence, the validity of the SFS should be tested also in patients with WAD in other health care systems. Secondly, in one study patients were classified as RTW if they had worked at least 1 day in the follow-up period [5]. These differences influence the proportion of patients classified as RTW or N-RTW, and therefore the results concerning the predictive properties of the SFS [33]. Third, the differences in symptoms of patients with WAD differ in part from those with CLBP. And forth, the depicted tasks of the SFS involving the spine may be perceived to the neck differently from the lower back.

Future studies should investigate whether a short version of the SFS would lead to similar measurement properties. Computer based measures could offer some advantages over a paper form. By using Item response theory (IRT) techniques only suitable items are assigned based on the response pattern of the evaluee. First results using a computer based measure similar to the SFS are promising, but need further evaluation in clinical samples [34,35].

Limitations

We used hypotheses and cut-off points based on the results of previous studies. These cut-offs may viewed as arbitrary. We used WC in % of the actual work. This may lead to differences in estimates of productivity loss compared to self-report of the employee, or other reporting measures [36-38]. Moreover, the alternative (WC) also has shortcomings; its psychometric properties are unknown and WC may rely on physicians interpretations and patients report [39]. Finally, replication studies are needed because the results differ in other populations, contexts and FCE procedures.

CONCLUSION

In patients with sub-acute WAD test-retest reliability, internal consistency, construct- and structural validity of the SFS were adequate. LoA was substantial. Sensitivity to accurately predict N-RTW was poor.

Based on the AUC the predictive validity of the SFS for N-RTW of patients with sub-acute WAD from an outpatient work rehabilitation setting was only sufficient for the short term (1 month).

ACKNOWLEDGEMENTS

The authors thank the physiotherapists and physicians of the Department of Work Rehabilitation, Rehaklinik Bellikon for their help in performing the tests and collecting data. We also thank Claudia Diethelm, Axel Gehrke and Stephan Scholz-Odermatt for data preparation, and all subjects for their participation.

REFERENCES

1. Burrus C, Ballabeni P, Deriaz O, Gobelet C, Luthi F. Predictors of nonresponse in a questionnaire-based outcome study of vocational rehabilitation patients. Arch Phys Med Rehabil. 2009;90:1499-505.
2. Matheson LN, Matheson ML, Grant J. Development of a measure of perceived functional ability. J Occup Rehabil. 1993;3:15-30.
3. Matheson LN. History, design characteristics, and uses of the pictorial activity and task sorts. J Occup Rehabil. 2004;14:175-95.
4. Oliveri M. Functional Capacity Evaluation. In Gobelet C, Franchignoni F eds. Vocational Rehabilitation. Paris: Springer, 2005.
5. Oesch PR, Hilfiker R, Kool JP, Bachmann S, Hagen KB. Perceived functional ability assessed with the spinal function sort: is it valid for European rehabilitation settings in patients with non-specific non-acute low back pain? Eur Spine J. 2010;19:1527-33.
6. Borloz S, Trippolini MA, Ballabeni P, Luthi F, Deriaz O. Cross-Cultural Adaptation, Reliability, Internal Consistency and Validation of the Spinal Function Sort (SFS) for French- and German-Speaking Patients with Back Complaints. J Occup Rehabil. 2012;22:387-93.
7. Robinson RC, Kishino N, Matheson L, Woods S, Hoffman K, Unterberg J, et al. Improvement in postoperative and nonoperative spinal patients on a self-report measure of disability: the Spinal Function Sort (SFS). J Occup Rehabil. 2003;13:107-13.
8. Gibson L, Strong J. The reliability and validity of a measure of perceived functional capacity for work in chronic back pain. J Occup Rehabil. 1996;6:159-75.
9. Suva. Suva: an overview [Swiss Accident Insurance Fund] 2013. Available from: http://www.suva.ch/english/startseite-en-suva/suva-en-suva/ueberblick-en-suva.htm. Accessed 17.09.2013.
10. U.S. Department of Labor. The Revised Handbook for Analyzing Jobs. 4th ed. Indianapolis: JIST Works, inc., 1991.
11. Stöckli H, Ettlin T, Gysi F, Knüsel O, Marelli R, Soltermann B. [Diagnostics and therapeutic approach in the chronic phase of whiplash associated disorders]. Schweiz Med Forum. 2005;5:1182-7.
12. Fitforwork-swiss. WOCADO [Workcapacity estimation for doctors] [The Work Foundation], 2013. Available from: http://www.fitforwork-swiss.ch/de/projekte.html. Accessed 03.12.2013.

13. Isernhagen SJ. Functional capacity evaluation: rational, procedure, utility of the kinesiophysical approach. J Occup Rehabil. 1992;2:157-68.

14. Trippolini MA, Reneman MF, Jansen B, Dijkstra PU, Geertzen JH. Reliability and safety of functional capacity evaluation in patients with whiplash associated disorders. J Occup Rehabil. 2013;23:381-90.

15. Ferraz MB, Quaresma MR, Aquino LR, Atra E, Tugwell P, Goldsmith CH. Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. J Rheumatol. 1990;17:1022-4.

16. Pool JJ, Ostelo RW, Hoving JL, Bouter LM, de Vet HC. Minimal clinically important change of the Neck Disability Index and the Numerical Rating Scale for patients with neck pain. Spine (Phila Pa 1976). 2007;32:3047-51.

17. MacDermid JC, Walton DM, Avery S, Blanchard A, Etruw E, McAlpine C, et al. Measurement properties of the neck disability index: a systematic review. J Orthop Sports Phys Ther. 2009;39:400-17.

18. Vernon H. The Neck Disability Index: state-of-the-art, 1991-2008. J Manipulative Physiol Ther. 2008;31:491-502.

19. Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale. An updated literature review. J Psychosom Res. 2002;52:69-77.

20. Herrmann C. International experiences with the Hospital Anxiety and Depression Scale--a review of validation data and clinical results. J Psychosom Res. 1997;42:17-41.

21. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res. 1995;4:293-307.

22. Portney LG, Watkins MP. Reliability. Foundations of clinical research. Applications to practice. 2nd ed. Upper Saddle River, NJ: Prentice-Hall Health, 2000:p.61-77.

23. Polit D, Beck C. Developing and testing self-report scales. In Polit D, Beck C eds. Nursing research, generating and assessing ecidence for nursing practice. Philadelphia: Lippincott, 2008:474-505.

24. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med. 1990;20:337-40.

25. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307-10.

26. Hazard Munro B. Statistical Methods for Health Care. Philadelphia: J. B. Lippincott, 1986.

27. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60:34-42.

28. de Vet HC, Terwee CB, Mokkink LB, Knol D. Measurement in Medicine: a practical guideed. Cambridge Cambridge University Press, 2011.

29. Streiner DL, Norman GR. Health measurement scales. A practical guide to their deverlopment and use. 4th ed. Oxford: Oxford University Press, 2008.

30. Carroll LJ, Hogg-Johnson S, Cote P, van der Velde G, Holm LW, Carragee EJ, et al. Course and prognostic factors for neck pain in workers: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine (Phila Pa 1976). 2008;33:S93-100.

31. Federal Supreme Court of Switzerland, 2010. Available from: http://www.bger.ch/index/juridiction/jurisdiction-inherit-template/jurisdiction-recht-leitentscheide1954-direct.htm. Accessed 07.03.2014, 2014.

32. Schrader H, Obelieniene D, Bovim G, Surkiene D, Mickeviciene D, Miseviciene I, et al. Natural evolution of late whiplash syndrome outside the medicolegal context. Lancet. 1996;347:1207-11.

33. Portney LG, Watkins MP. Validity. Foundations of clinical research. Applications to practice. 2nd ed. Upper Saddle River, NJ: Prentice-Hall Health, 2000:p.79-107.

34. Mayer J, Mooney V, Matheson L, Leggett S, Verna J, Balourdas G, et al. Reliability and validity of a new computer-administered pictorial activity and task sort. J Occup Rehabil. 2005;15:203-13.

35. Mooney V, Matheson LN, Verna J, Leggett S, Dreisinger TE, Mayer JM. Performance-integrated self-report measurement of physical ability. Spine J. 2010;10:433-40.

36. Gatchel RJ. Psychosocial factors that can influence the self-assessment of function. J Occup Rehabil. 2004;14:197-206.

37. Tompa E. Measuring the burden of work disability: a review of methods, measurement issues and evidence. In Loisel P, Anema JR eds. Handbook of Work Disability. Prevention and Management. New York: Springer, 2013:43-58.

38. Gauthier N, Sullivan MJ, Adams H, Stanish WD, Thibault P. Investigating risk factors for chronicity: the importance of distinguishing between return-to-work status and self-report measures of disability. J Occup Environ Med. 2006;48:312-8.

39. Rainville J, Pransky G, Indahl A, Mayer EK. The physician as disability advisor for patients with musculoskeletal complaints. Spine (Phila Pa 1976). 2005;30:2579-84.

Chapter 8

Cross-cultural adaptation, reliability, internal consistency and validation of the Spinal Function Sort (SFS) for French- and German-speaking patients with back complaints

Stephane Borloz
Maurizio A. Trippolini
Pierluigi Ballabeni
Francoise Luthi
Oiliver Deriaz

Journal of Occupational Rehabilitation 2012;22:387-393.

ABSTRACT

Introduction: Functional subjective evaluation through questionnaire is fundamental, but not often realized in patients with back complaints, lacking validated tools. The Spinal Function Sort (SFS) was only validated in English. We aimed to translate, adapt and validate the French (SFS-F) and German (SFS-G) versions of the SFS.

Methods: 344 patients, experiencing various back complaints, were recruited in a French (n=87) and a German-speaking (n=257) center. Construct validity was estimated via correlations with SF-36 physical and mental scales, Pain Intensity and Hospital Anxiety and Depression Scales (HADS). Scale homogeneities were assessed by Cronbach's α. Test-retest reliability was assessed on 65 additional patients using intraclass correlation (IC).

Results: For the French and German translations, respectively, α were 0.98 and 0.98; IC 0.98 (95% CI: [0.97; 1.00]) and 0.94 [0.90; 0.98]. Correlations with physical functioning were 0.63 [0.48; 0.74] and 0.67 [0.59; 0.73]; with physical summary 0.60 [0.44; 0.72] and 0.52 [0.43; 0.61]; with pain -0.33 [-0.51; -0.13] and -0.51 [-0.60; -0.42]; with mental health -0.08 [-0.29; 0.14] and 0.25 [0.13; 0.36]; with mental summary 0.01 [-0.21; 0.23] and 0.28 [0.16; 0.39]; with depression -0.26 [-0.45; -0.05] and -0.42 [-0.52; -0.32]; with anxiety -0.17 [-0.37; -0.04] and -0.45 [-0.54; -0.35].

Conclusions: Reliability was excellent for both languages. Convergent validity was good with SF-36 physical scales, moderate with VAS pain. Divergent validity was low with SF-36 mental scales in both translated versions and with HADS for the SFS-F (moderate in SFS-G). Both versions seem to be valid and reliable for evaluating perceived functional capacity in patients with back complaints.

INTRODUCTION

The follow-up of patients with musculoskeletal disorders in clinical and research settings is not only based on clinical exams or radiography but also on self-administered questionnaires which are inexpensive and give insight into the patient's perspective.

In occupational rehabilitation, one important activity is the functional capacity evaluation (FCE) of patients in order to determine readiness or ability for safe return to work following musculoskeletal injury [14]. The patient's self efficacy (SE) level was proposed as a relevant psychosocial factor that may influence FCE. Perceived SE refers to the individual's beliefs about their own competence or ability [2]. SE beliefs may influence the patient's behavior, e.g. the ability to overcome negative experiences. It has been suggested that SE is more closely related to work disability than actual physical abilities [34]. Assessment of SE by self-report therefore plays an important role in predicting health outcome [19,22]. It has also been recommended that patients with low back pain should be assessed with both instruments (i.e. self-report and performance tests) because these strategies may lead to different results [23,36,37].

Self-administered questionnaires should be developed with accurate and rigorous instruments to ensure that they are specific to the studied concept as well as reliable and responsive (clinimetric qualities) [11,28]. A great variety of questionnaires have been developed to assess the perceived function of patients with back pain. Some of them such as the Oswestry Disability Index, the Roland Morris Disability Questionnaire and the Quebec Back Pain Disability Scale have been recommended for clinical purposes by an expert panel [13]. The utility of questionnaires in rehabilitation settings is often limited by the literacy level of the patients [16]. One approach to improving the comprehension of the questionnaire by patients with low literacy levels is to inform the patient through pictorial activities and task sorts (PATS) designed for self-assessment of functional ability in occupational rehabilitation such as were developed in the seventies [24].

Recently, efforts have focused on the creation of questionnaires more oriented towards functional limitations and occupational perspectives [27]. However, those picture-based questionnaires are often validated in English only, and not for use by non-English speaking patients.

The Spinal Function Sort (SFS), published in English in 1989 [25], has proven to be of advantage in work-related rehabilitation settings [18,20,21,35,39]. It is often used in addition to functional capacity evaluations to assess the self-perceived functional capacity of patients with back complaints [30]. It is a picture-based generic tool that is useful for all kinds of back disorders. The reliability and validity of the SFS have been reported [15,26,29] but, to the best of our

knowledge, no German or French versions have been properly cross-culturally adapted and translated. The aim of this study was to do a cross-cultural adaptation and validation of the SFS in French and German.

METHODS

Spinal Function Sort (SFS)

The French and German translations of the Spinal Function Sort consists, as the original SFS, of a booklet containing drawings (Figure 8.1) with a brief a description of 50 tasks. These tasks are performed by men and women and reflect a wide range of daily living or vocational activities that involve the spine. The pictured activities are graded from light to heavy material handling, so that scores can be compared to the Physical Demand Characteristics from the United States Department of Labor's Dictionary of Occupational Titles [1]. Subjects are asked to answer quickly without spending too much time on any one drawing. They are told that their "first impression is usually the best". There is no time limit to fill out the questionnaire. Subjects rate their ability to perform the task on a 5-point Lickert scale (from "able" to "restricted" to "unable"). An additional category depicted as "?" means "I don't know", for example, for an unfamiliar task. Items are scored from 4 (able) to 0 (unable or "?"). The SFS is scored manually by the assessor and yields a total score, which can range from 0 to 200. This total score corresponds to the level of perceived physical work, ranging from sedentary to very heavy, and can be compared to the Dictionary of Occupational Titles.

Figure 8.1 Item 27 of the Spinal Function Sort (SFS) questionnaire: Load or unload a dishwasher.

Following the scoring instruction of the original SFS, questionnaires with 4 or more "I don't know" responses, were excluded from the present study because of potential bias. Moreover, the SFS has 2 internal validity check drawings with the same questions but different images to test the reliability of subjects (questions #6 and #50; #17 and #49). Subjects who showed inconsistencies greater than 3 points on the 5 point scale were also excluded.

Cross-cultural adaptation

The cross-cultural adaptation of the SFS was performed according to the guidelines of the American Academy of Orthopedic Surgeons (AAOS) Outcomes Committee [3] and as recommended by others in the literature [17,41]. The following five steps were documented in a written report: (1) Forward translation from English to French and to German by two translators whose native language was French, or German, and fluent in English (T1 and T2). One of the translators was informed about the aims of the study, and the other received only limited information (so-called naïve translator). Moreover, none of the translators were physicians. (2) Synthesis of T1 and T2 were amalgamated to form the unique translated version T12 by resolving any discrepancies under supervision of a methodologist who was not involved in the translation process. (3) Back translation of the T12 version from French or German into English by two translators whose native language was English, and who were fluent in French, or German (BT1 and BT2). These two translators were naïve to the study and not directly linked with the medical domain. (4) Consensus meeting with all the involved subjects (translators, methodologist, specialist physicians in occupational rehabilitation) in order to resolve any discrepancies and doubts met during the translation, and to establish the pre-final French and German versions of the SFS. (5) Pre-testing of the French and German versions for the accuracy of the words and ease of understanding of the SFS was conducted with 20 consecutive patients with back complaints. Patients were asked to mention any difficulties encountered during a phone call. The last steps were realized by submitting the final version of the French SFS (SFS-F) and German SFS (SFS-G) and all reports and forms to a committee keeping track of the translated version in order to verify that the recommended stages were followed.

Participants

For each language, two sets of participants were recruited: one for the assessment of construct validity and scale homogeneity, and a second set for test-retest reliability.

For construct validity of the French version, 17 women and 70 men were recruited. These 87 subjects were consecutive inpatients hospitalized because of persistent back pain between

2004 and 2005 at the Clinique romande de réadaptation at Sion, Switzerland. The mean age was 44 years (SD: 10; range: 19 to 61). Test-retest reliability of the French version was assessed on a sample of 21 patients (9 women, 12 men; mean age 43 years, SD: 14, range 19 to 65) recruited in 2009. In addition to a history of back pain, subjects in both samples had diagnoses such as fracture (operated or treated conservatively), discal prolapse or hernia, degenerative disorders, discopathies, status after discal hernia operation, contusion(s), tight canal, olisthesis, spina bifida occulta, isthmic lysis, non specific lumbalgia, whiplash, cervical strain, transitional anomaly, Scheuermann's disease.

Construct validity of the German version was assessed on 257 consecutive inpatients hospitalized between November 2003 and February 2006 (53 women, 204 men; mean age 40 years, SD 11, range 18 to 64). These subjects were recruited at the Rehaklinik Bellikon in Bellikon, Switzerland, because of persistent back pain. Test-retest reliability of the German version was assessed on a convenience sample of 51 patients (9 women; 41 men, mean age 43.6 years, SD 13 years, range 21 to 65) recruited in 2009. Diagnoses for both samples were similar to the French cohort.

Patients with upper and/or lower limb complaints were excluded because of the risk of influencing the SFS scores. Patients with psychopathology in which pain is the central element (such as somatoform trouble) were also excluded. Patients who had other non-disabling psychopathologies were included.

The study was approved by the ethical committees of the canton Valais and the canton Aargau, where the two clinics are located. All patients signed a written informed consent form.

Validation

All patients completed the French or German version of the SFS, the Medical Outcomes Short Form (SF-36) [40], the Hospital Anxiety and Depression Scale (HADS) [5], and the Visual Analogue Scale for Pain Intensity (VAS) [8]. Construct validity of the SFS translated versions was assessed by estimating Pearson's correlation coefficients between the French (resp. German) versions of the SFS and the HADS, VAS, and relevant subscales of the SF-36. The Physical Functioning subscale (PF), the Physical Summary Scale (PCS) and the VAS were used to assess convergent validity (high correlations expected); the Mental Health scale (MH), the Mental Summary Scale (MCS) and HADS for divergent validity (low correlations expected). Ninety-five percent confidence intervals for the correlation coefficients were calculated by means of Fisher's transformation.

Ceiling and floor effects were defined as present if at least 15% of results reached the maximum or the minimum value [4].

Internal consistency was determined by Cronbach's α [10,31], which is a general coefficient of homogeneity between items. Values for α can range from 0 (no internal consistency) to 1 (perfect internal consistency), where a value above 0.8 is considered acceptable [33].

The reliability of the translated versions was assessed by test-retest reliability and quantified by the intraclass correlation coefficient (ICC) [6]. Patients (see above) who were not expected to have a significant health status change between tests were asked to fill out the SFS-F (resp. SFS-G) on two occasions separated by two days. Values for ICC can range from 0 (no agreement) to 1 (perfect agreement). Bland-Altman plots were used to assess the disagreement between test and retest values [7]. Such plots show the individual score differences between tests as a function of the individual mean scores of the two tests. 95% limits of agreement were calculated as the mean difference ± 1.96 SD of the difference. The narrower the limits of agreement, the smaller the disagreement between the repeated tests.

All calculations were performed using the statistical package Stata 11.0 for Windows (StataCorp LP, 4905 Lakeway Drive, College Station, TX 77845, USA).

RESULTS

Cross cultural adaptation

The translations and back-translations of the SFS items were carried out in both French and German without any relevant difficulties. The back-translations of the T12 versions to English were very similar to the original versions. Only some typically US expressions or words were different as our back-translators were native from the United Kingdom and India. Moreover, patients did not mention any difficulties in understanding the items.

Validation

SFS-French version

Eighty-seven subjects were eligible for the validation of the SFS-F. The excluded patients had inconsistency in the internal validity check or more than 4 "I don't know" answers.

For convergent validity, we found a correlation coefficient of 0.63 (95% CI: 0.48 to 0.74) between SFS-F and PF, 0.60 (95% CI: 0.44 to 0.72) between SFS-S and PSC, and -0.33 (95% CI: -0.51 to -0.13) between SFS-F and VAS. The assessment of divergent validity resulted in an SFS-F-MH correlation of -0.08 (95% CI: -0.29 to 0.14), an SFS-F-MCS correlation of 0.01 (95% CI: -0.21 to 0.23), an SFS-F-HADS depression correlation of -0.26 (95% CI: -0.45 to -0.05), and an SFS-F- HADS anxiety correlation of -0.17(95% CI: -0.37 to -0.04) (Table 8.1).

No evidence for floor or ceiling effects was found for the total score since no patient reached the minimum or maximum possible score. A floor effect was found in the items 45–48 with more than 99 % of the participants rating their ability to perform the task as "restricted" (14%) or "unable" (85%) on a 5-point Lickert scale from "able" to "restricted" to "unable"). For internal consistency, Cronbach's α was 0.98 for the SFS-F. The reliability, assessed by test-retest in 21 patients, resulted in ICC values of 0.98 (95% CI: 0.97 to 1.00). The mean difference between test and retest was 0.3, with 95% upper and lower limits of agreement at -11.5 and 12.1 (Figure 8.2).

SFS-German version

Three hundred and nine subjects were eligible for validation of the SFS-G. The excluded patients had inconsistency in the internal validity check or more than 4 "I don't know" answers.

For convergent validity, we found a correlation coefficient of 0.67 (95% CI: 0.59 to 0.73) between SFS-G and PF, 0.52 (95% CI: 0.43 to 0.61) between SFS-G and PSC, and -0.51 (95%

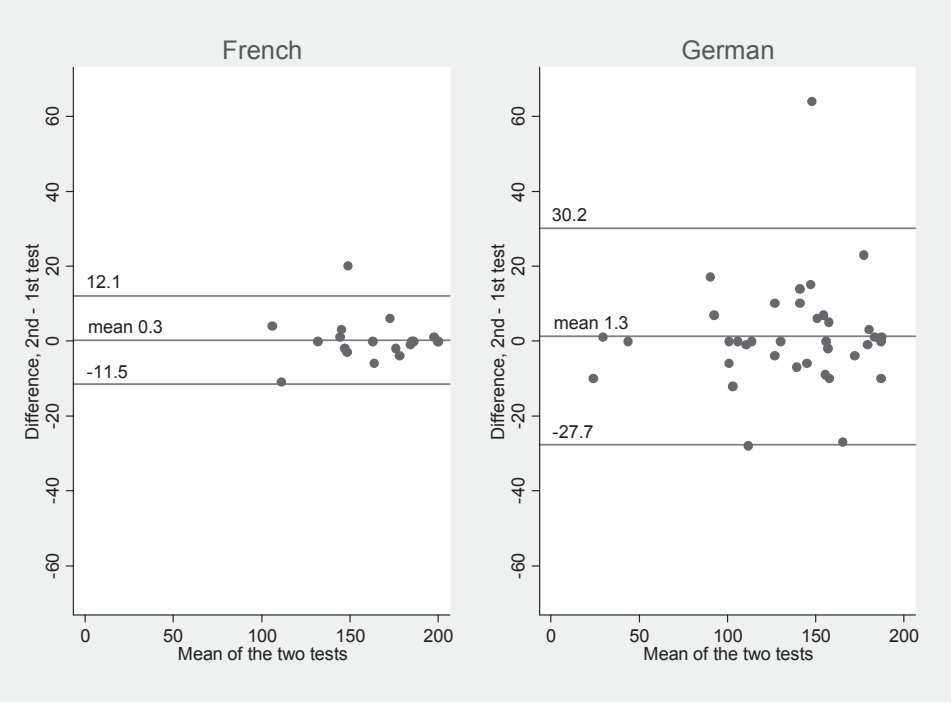


Figure 8.2 Bland-Altman plot for the two languages. The middle line represents the mean difference between the two tests. The upper and lower lines represent the upper and lower limits of agreement, i.e. mean difference + 1.96 SD of the differences and mean difference - 1.96 SD of the differences, respectively.

CI: -0.60 to -0.42) between SFS-G and VAS. The assessment of divergent validity resulted in an SFS-G-MH correlation of 0.25 (95% CI: 0.13 to 0.36), an SFS-G-MCS correlation of 0.28 (95% CI: 0.16 to 0.39), an SFS-G-HADS depression correlation of -0.42 (95% CI: -0.52 to -0.32), and an SFS-G- HADS anxiety correlation of -0.45 (95% CI: -0.54 to -0.35).

No evidence for floor or ceiling effects was found for the total score since no patient reached the minimum possible score and only one scored the maximum possible value. A floor effect was found in items 45–48 with more than 97% of the participants rating their ability to perform the task as "restricted" (12%) to "unable" (85%). For internal consistency, Cronbach's α was 0.98 for the SFS-G. Reliability, assessed by test-retest in 44 patients, resulted in ICC values of 0.94 (95% CI: 0.90 to 0.98). Mean difference between test and retest was 1.3, with 95% lower and upper limits of agreement at -27.7 and 30.2 (Figure 8.2). A look at Figure 8.2 shows a highly influential patient with a difference of over 60 units between tests. Limits of agreement calculated without that patient were -20.9 and 19.9 for a mean difference of -0.5.

DISCUSSION

The original English version of the SFS was translated and adapted into French and German, respectively, to create the SFS-F and SFS-G versions. Evidence for reliability and validity was shown, supporting the use of the SFS-F and SFS-G as a self-report instrument for individuals with a wide range of chronic back disorders. Specifically, evidence for convergent validity, divergent validity, internal structure, and score stability were provided for the SFS-F and SFS-G.

Typical activities, especially those regarding gardening with specific tools, had to be adapted for the French and German culture. For example, a "spade-shovel" is not commonly used by patients in our countries and was modified as "shovel" ("pelle" in French and "Schaufel" in German). Thus, and although the SFS is a pictorial questionnaire, cross-cultural adaptation shows the importance of following the complete AAOS guidelines for a valuable final version.

As hypothesized for convergent validity, the correlation coefficients between the translated SFS versions and the SF-36 physical scales were fairly high, i.e. 0.63 and 0.67 for the SFS-F and SFS-G, respectively. Moreover, they were similar to values found by Gibson et al. (1996) with other scales such as the Pain Disability Index (-0.64), the Work Reentry Questionnaire (0.67), the Self Efficacy Questionnaire (0.55) and the Pain Self-Efficacy Questionnaire (0.78) [15]. Those questionnaires could not have been used in the present study because of the lack of French and German validated versions. The pain scale also showed a significant correlation with the SFS-G (-0.51), but only a low correlation with the SFS-F (-0.33). This last estimation is rather imprecise (95% CI: -0.51 to -0.13), probably due to a smaller sample compared to the German version. For divergent validity, we found no correlation (-0.08) between the SF-

36 mental scales and the SFS-F, and a low correlation with the SFS-G (0.25) as hypothesized. The small differences between the French and German versions might be explained either by some cultural differences regarding the implication of back problems in daily living and, consequently, the interaction with mental health of the SF-36 (which has questions regarding irritability, sadness, motivation), or by sampling. The correlation between HADS and the SFS was low (0.26) for the French version and moderate (0.42) for the German version. These correlations are possibly due to the chronicity of back problems in our patients, who were recruited in tertiary centers. Patient populations with chronic occupational back pain are known to exhibit higher prevalence of psychological disorders compared to the general population [12]. Moreover, the difference between the French and German versions may, as for mental health, be explained by cultural differences related to either the patients or medical practice, but also to the difference in the timing of hospitalization in the two centers after back problems were diagnosed. Furthermore, it must be kept in mind that our study samples were not randomly drawn from a population but were convenience samples. A previous study performed at the Clinique Romande de réadaptation Suva care (Switzerland) has shown that questionnaire responders differed from non-responders in some sociodemographic and biophsychosocial aspects [9]. Thus, some degree of selection bias, which may differ between clinics, may well have occurred in the present study.

A floor effect was found in items 45–48. Those items describe activities where weights of 50 kg are lifted either from floor, waist or overhead height or down again. Most participants felt they could not carry out such strenuous activities. It may be questioned whether these items are of great value for the clinical purpose of the questionnaire. Furthermore, lifting tasks involving weights over 25 kg are nowadays prohibited in most occupations in Switzerland, France and Germany.

According to the literature, a Cronbach's α over 0.80 (over 0.90 for clinical applications) represents a good internal consistency. We found excellent α values far above these thresholds with 0.98 for both SFS-F and SFS-G. This high internal consistency may be partly influenced by the high number of items since α has the property of becoming larger with increasing item number, given equal between-item correlations [38]. However, our values are similar to those of the English versions (0.98), suggesting that the French and German translations bear the same level of internal consistency as the original version.

The reliability of both the SFS-F and the SFS-G was excellent with regard of an ICC of 0.98 and 0.94, respectively. These coefficients are higher than the values reported in the original version (0.89) [26]. Moreover, the confidence intervals (0.97 to 1.00 for the French version, and 0.90 to 0.98 for the German version) were narrow for both translations, indicating rather precise estimates.

The limits of agreement were calculated to determine the magnitude of disagreement between the two measurement occasions. With all patients included, the interval between the limits of agreement of the German version was over twice that of the French version (57.9 and 23.6 units, respectively). After exclusion of a highly influential patient, the German version's interval was reduced to 41.8 units. However, further studies should be done to evaluate the minimal clinical important change [32] to establish whether the difference in score is clinically relevant.

Some limitations of the present study have to be recognised. First, only patients hospitalized in tertiary centers for chronic back problems were included. Thus, results of SFS-F and SFS-G questionnaires have to be interpreted with caution in other clinical settings. The use of convenience samples instead of random samples was discussed above.

Although the SFS has been successfully used for the last 20 years, some recommendations may be given here to improve its clinical utility. First, old fashioned drawings (i.e. old type of vacuum cleaner) should be replaced by new pictures of tools used nowadays. Second, reduction in the number of items would lead to important time saving and therefore further improve its clinical utility. In this context, we calculated PACT scores using either the 25 even or the 25 uneven items and ranked the patients on the full score and the two half-scores. Correlations between the full score rank and each of the half-score ranks were 0.99 for both languages, showing item redundancy. A reduction in the number of items is also supported by the high internal consistency. Third, relevant items which include posture of spinal load, such as sitting, should be included in the SFS. The development of a brief version of the SFS is clearly needed. Further studies exploring these measurement properties in different settings and with other validation tools are therefore needed.

In conclusion, the French and the German versions of the Spinal Function Sort, seem to be valid and reliable, and it is a tool that is easy to administer to evaluate perceived functional capacity for native French-speaking and German-speaking patients with back disorders, both for clinical purposes and research.

ACKNOWLEDGEMENTS

This study was supported by Clinique romande de réadaptation and the Rehaklinik Bellikon. The authors thank Fabienne Reynard, Dominique Buchard, Mike Short, Trevor Mc Intosh, Amisha Gudibanda, Marilyn Murbach, Andrea Müller-Hildebrand, Julia Koelle, and Peter Erhart for their translation and back-translation work, and all patients for their participation. We also thank the physiotherapists for their help in recruiting patients for this study, and Peter Erhart for his essential support with data storage, involvement in the translation process and his role as guarantor of the study at the Rehaklinik Bellikon.

REFERENCES

1. The Revised Handbook for Analyzing Jobs. Washington, DC: US Department of Labor, Employment and Training Administration, 1991.
2. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev. 1977;84:191-215.
3. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1976). 2000;25:3186-91.
4. Bent NP, Wright CC, Rushton AB, Batt ME. Selecting outcome measures in sports medicine: a guide for practitioners using the example of anterior cruciate ligament rehabilitation. Br J Sports Med. 2009;43:1006-12.
5. Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale. An updated literature review. J Psychosom Res. 2002;52:69-77.
6. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med. 1990;20:337-40.
7. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307-10.
8. Boonstra AM, Schiphorst Preuper HR, Reneman MF, Posthumus JB, Stewart RE. Reliability and validity of the visual analogue scale for disability in patients with chronic musculoskeletal pain. Int J Rehabil Res. 2008;31:165-9.
9. Burrus C, Ballabeni P, Deriaz O, Gobelet C, Luthi F. Predictors of nonresponse in a questionnaire-based outcome study of vocational rehabilitation patients. Arch Phys Med Rehabil. 2009;90:1499-505.
10. Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297-334.
11. de Vet HC, Terwee CB, Bouter LM. Current challenges in clinimetrics. J Clin Epidemiol. 2003;56:1137-41.
12. Dersh J, Gatchel RJ, Mayer T, Polatin P, Temple OR. Prevalence of psychiatric disorders in patients with chronic disabling occupational spinal disorders. Spine (Phila Pa 1976). 2006;31:1156-62.
13. Deyo RA, Battie M, Beurskens AJ, Bombardier C, Croft P, Koes B, et al. Outcome measures for low back pain research. A proposal for standardized use. Spine (Phila Pa 1976). 1998;23:2003-13.
14. Genovese E, Galper JS. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations: American Medical Association, 2009.
15. Gibson L, Strong J. Assessment of Psychosocial Factors in Functional Capacity Evaluation of Clients with Chronic Back Pain. British Journal of Occupational Therapy. 1998;61:399-404.
16. Gonzalez-Calvo J, Gonzalez VM, Lorig K. Cultural diversity issues in the development of valid and reliable measures of health status. Arthritis Care Res. 1997;10:448-56.
17. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol. 1993;46:1417-32.

18. Henchoz Y, de Goumoens P, So AK, Paillex R. Functional multidisciplinary rehabilitation versus outpatient physiotherapy for non specific low back pain: randomized controlled trial. Swiss Med Wkly. 2010;140:131-3.

19. Holden G. The relationship of self-efficacy appraisals to subsequent health related outcomes: a meta-analysis. Soc Work Health Care. 1991;16:53-93.

20. Innes E, Hardwick M. Actual versus perceived lifting ability in healthy young men (18-25 years). Work. 2010;36:157-66.

21. Kool JP, Oesch PR, Bachmann S, Knuesel O, Dierkes JG, Russo M, et al. Increasing days at work using function-centered rehabilitation in nonacute nonspecific low back pain: a randomized controlled trial. Arch Phys Med Rehabil. 2005;86:857-64.

22. Lackner JM, Carosella AM. The relative influence of perceived pain control, anxiety, and functional self efficacy on spinal function among patients with chronic low back pain. Spine (Phila Pa 1976). 1999;24:2254-60; discussion 60-1.

23. Lee CE, Simmonds MJ, Novy DM, Jones S. Self-reports and clinician-measured physical function among patients with low back pain: a comparison. Arch Phys Med Rehabil. 2001;82:227-31.

24. Matheson LN. History, design characteristics, and uses of the pictorial activity and task sorts. J Occup Rehabil. 2004;14:175-95.

25. Matheson LN, Matheson M. Spinal Function Sort. Rating of Perceived Capacity. Test Booklet and Examiners Manualed. Trabuco Canyon, California: Performance and Capacity Testing, 1989.

26. Matheson LN, Matheson ML, Grant J. Development of a measure of perceived functional ability. Journal of Occupational Rehabilitation. 1993;3:15-30.

27. Mayer J, Mooney V, Matheson L, Leggett S, Verna J, Balourdas G, et al. Reliability and validity of a new computer-administered pictorial activity and task sort. J Occup Rehabil. 2005;15:203-13.

28. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63:737-45.

29. Oesch PR, Hilfiker R, Kool JP, Bachmann S, Hagen KB. Perceived functional ability assessed with the spinal function sort: is it valid for European rehabilitation settings in patients with non-specific non-acute low back pain? Eur Spine J. 2010;19:1527-33.

30. Oliveri M. Functional Capacity Evaluation. In Gobelet C, Franchignoni F eds. Vocational Rehabilitation. Paris: Springer, 2005.

31. Osburn HG. Coefficient alpha and related internal consistency reliability coefficients. Psychol Methods. 2000;5:343-55.

32. Ostelo RW, Deyo RA, Stratford P, Waddell G, Croft P, Von Korff M, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. Spine (Phila Pa 1976). 2008;33:90-4.

33. Portney LG, Watkins MP. Foundations of Clinical Research. Application to Practice. 3 ed: Prentice-Hall, 2007.

34. Reneman MF, Jorritsma W, Schellekens JM, Goeken LN. Concurrent validity of questionnaire and performance-based disability measurements in patients with chronic nonspecific low back pain. J Occup Rehabil. 2002;12:119-29.

35. Robinson RC, Kishino N, Matheson L, Woods S, Hoffman K, Unterberg J, et al. Improvement in postoperative and nonoperative spinal patients on a self-report measure of disability: the Spinal Function Sort (SFS). J Occup Rehabil. 2003;13:107-13.

36. Schiphorst Preuper HR, Reneman MF, Boonstra AM, Dijkstra PU, Versteegen GJ, Geertzen JH, et al. Relationship between psychological factors and performance-based and self-reported disability in chronic low back pain. Eur Spine J. 2008;17:1448-56.

37. Smeets RJ, van Geel AC, Kester AD, Knottnerus JA. Physical capacity tasks in chronic low back pain: what is the contributing role of cardiovascular capacity, pain and psychological factors? Disabil Rehabil. 2007;29:577-86.

38. Streiner DL, Norman GR. Health measurement scales. A practical guide to their deverlopment and use. 3rd ed. Oxford: Oxford University Press, 2003.

39. Sufka A, Hauger B, Trenary M, Bishop B, Hagen A, Lozon R, et al. Centralization of low back pain and perceived functional outcome. J Orthop Sports Phys Ther. 1998;27:205-12.

40. Ware JE, Jr., Gandek B. Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. J Clin Epidemiol. 1998;51:903-12.

41. Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, et al. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health. 2005;8:94-104.

Chapter 9

General discussion

INTRODUCTION

Whiplash-associated disorders (WAD) grades I and II cannot be diagnosed with common medical diagnostic tests such as imaging techniques e.g. MRI [1-4]. Moreover, as with many pain disorders, the casual pathway from biomedical findings to (work-related) disability is not evident [5,6]. Therefore, in patients with WAD, clinicians must rely on the information gathered with self-report measures, clinical assessments and performance measures such as Functional Capacity Evaluation (FCE) tests. The initiative of the Neck Pain Task Force and its associated Disorders (NPTF) concluded that "there is a need to establish reliability, validity, and utility of functional capacity testing" [7] [p. S118]. In addition the NPTF suggested "there is a need to measure all performance parameters simultaneously: reliability, validity, responsiveness to change, and easy of administration, in self-assessment questionnaires" [7] [p. S119]. Furthermore, there is ample room for methodological improvement of studies on FCE [8,9]

Given this background, the main focus of this thesis was to conduct an extensive evaluation of measures of functioning in patients with WAD. Seven studies have been performed. This chapter summarizes the main findings of this thesis; the methodological considerations, including the strengths and the limitations of studies are discussed, followed by the implications for raters, patients and referrers. Recommendations for future research are given. In the final part the valorization of the findings of this thesis are reported. This general discussion ends with final conclusions.

MAIN FINDINGS

In a systematic review on the validity of instruments that claim to detect submaximal capacity when maximal capacity is requested in patients with chronic nonspecific musculoskeletal pain evidence was found that submaximal capacity can be detected with using a lumbar motion monitor or visual observations alongside a FCE lifting test. However, this evidence was based on a small number of studies. Major advances are needed to enable the conduct of well-designed diagnostic studies using practical instruments in large clinical samples (*Chapter 2*).

The reliability of observational criteria to determine physical effort during FCE tests applied by clinicians is on average acceptable for material handling tests, but poor in postural tolerance and ambulation tasks. A dichotomous scale (patient does or does not perform with maximal effort) had higher intra- and inter-rater reliability than a four-point scale (light-medium, heavy, maximal or over maximal effort). The average reliability of a clinical sample of raters improved over a 10 month period, indicating that experience plays a role in the reliability of assessment, and therefore regular training may enhance the rater reliability (*Chapter 3*).

The test-retest reliability of the FCE tests in patients with WAD was good, based on a cut-off value of ICC >0.75 in 10 out of 12 FCE tests. Safety requirements were fulfilled. No serious adverse side-effects of the FCE tests were reported and within few days increased pain levels had returned to baseline levels. Results at the patient level should be interpreted with care because limits of agreement were substantial (*Chapter 4*).

The construct validity of FCE tests in a sample of patients with diverse cultural backgrounds with WAD was good. The majority of hypotheses about differences in gender, cultural groups and reference measures were not rejected (*Chapter 5*).

The FCE tests are not independent predictors of future work capacity in patients with WAD. A predictive model was developed. Work capacity at baseline, cultural background, self-reported disability and time course were independent predictors of future work capacity in patients with WAD (*Chapter 6*).

The Spinal Function Sort (SFS), a picture-based questionnaire, has good test-reliability and is valid for the measurement of perceived functional ability in patients with WAD with diverse cultural backgrounds. However, care should be taken in clinical interpretation of data at the individual patient level, because of the large limits of agreement. The predictive validity of the SFS in patients with sub-acute WAD as a means of determining future work capacity was only sufficient for short term (*Chapter 7*).

German and French versions of the SFS were cross-culturally adapted and translated. Both language versions have good test-retest reliability and good construct validity when compared with quality of life and mental distress scales. Care should be taken in clinical interpretation at the individual patient level, because of the large limits of agreement and considerable variability between the language groups i.e. French- and German-speaking groups (*Chapter 8*).

METHODOLOGICAL CONSIDERATIONS

Each of the studies in this thesis is subject to some limitations, which have been reported in the individual chapters. Reviewing the chapters the following general strengths and limitations should be considered.

STRENGTHS AND LIMITATIONS

Strengths

- This thesis was the first evaluation of the measurement properties of FCE tests in patients with WAD. Current standards for evaluation of health measurement instruments were used: setting *a priori* hypotheses, blinding evaluators and patients where appropriate, reporting measurement error on an individual level, calculating effect sizes for strength of associations, using larger sample sizes and applying longitudinal mixed models corrected for confounders in prognostic studies [10,11].
- Known biological, psychological and social determinants of WAD were investigated in the research reported in this thesis. Specifically, self-report measures, performance tests, and social variables have been analyzed.
- In much health care research 'hard-to-reach' patients are underrepresented [12-14]. Participants with minority cultural backgrounds were adequately represented in this thesis.
- The FCE tests applied in this thesis were embedded in the patients' usual care setting. It is particularly important that the measures used in this study do not require expensive equipment or time consuming analysis [15]. Raters who participated in this study were taken from a representative sample of clinicians. Taking these factors into account adds to the generalizability of the results in the Swiss health care system.
- A reliable submaximal effort score for FCE tests was developed.
- As part of the research for this thesis replication of studies previously performed only by few research groups in Canada and the Netherlands was performed. Replication may be important, because studies of FCE tests performed in different countries with different jurisdictions, healthcare and social-security systems have produced variable results [16,17].

Limitations

- Generalizability of results in this thesis may be limited due to use of a selected sample of patients, characteristics of the evaluation including the tests and the raters, specific jurisdiction, other influencing external factors.
- Psychological variables not included in this thesis, such as perceived injustice and post-traumatic-stress beliefs [18], may be relevant in WAD.
- Self-reported impaired health pre-injury is strongly associated with reporting WAD [19]. This thesis included data on the number of injuries prior to WAD, but the general health

and neck pain of the participants were unknown. In the prognostic study the treating General Practitioner (GP), case manager, physiotherapist and occupational physician had access to the results of the FCE tests and these may have influenced their rating. There was no control for co-interventions during the period 6 to 52 weeks after FCE, or for the type of work, which may be an important confounder for return to work (RTW) and work capacity (WC).

- RTW may be measured in different ways including self-report measures, derivation from social security system or employer administrative data on sick leave [20,21]. Large differences between types of RTW measures have been reported [22]. There is as yet no consensus on the best way of measuring RTW [23]. In this thesis, WC was used as a measure of ability to work. WC was determined by the treating physician, usually a general practitioner, and corresponds to the proportion of his or her pre-injury work activities of which an individual is capable. WC is directly related to compensation. Although this way of measuring RTW had some advantages it also has shortcomings; its psychometric properties are unknown and WC is often based on the patient's report and physician's interpretation [24]. It is suggested that the choice of the appropriate measure of RTW should be based on the question at issue, for example a self-report measure may be appropriate for a quality of life study, whereas for a cost-effectiveness study the number of days for which compensation is paid, may be relevant.

Other considerations

- Based on the inclusion criteria one might infer that the results of this thesis are only applicable to patients 'diagnosed' with WAD grade I or II according to the Québec Task Force (QTF) classification. WAD is not a disease, but a functional disorder; these patients share many symptoms and behaviors with other chronic disorders [25]. One might therefore conclude that the results of these studies are also applicable to other chronic conditions. Or in Dennis Turk's words: "…those who suffer from different conditions may have more in common than those with the same diagnosis" [26].
- FCE tests are influenced by various psychosocial factors and should always be interpreted in the biomedical and the contextual situation of the patient [27]. Hence, a FCE test score represents the ability of an individual at a given point in time, under certain circumstances [28-30]. This means that the results of FCE tests, self-report measures of function and workplace assessments may differ [31-33].
- Comprehensive self-report indices of work functioning which claim to measure perceived difficulties in meeting work demands of workers given their physical health or

emotional problems have been developed [34]. These indices consist of scales for work scheduling, work output demands and physical, mental and social demands [35,36]. FCE providers, employers and RTW specialists may consider these indices within the FCE format (*Introduction*, Table 1.2) although their measurement properties are still under investigation. Combining data from FCE tests, self-reported work functioning and individual workplace assessment may improve the quality of work-related evaluations and enhance RTW interventions directed at workers.

- The generic FCE tests used in this thesis are highly standardized and replicable, so the data thus obtained can be compared with other populations. One may argue that on an individual level a generic approach is less suited to determining work capacity than job-specific or onsite FCEs or individually tailored workplace assessments [8] but these methods have yet to be studied in patients with WAD.
- In Switzerland the measures used in this thesis, i.e. FCE tests and the SFS, are also recommended for the determination of WC in patients with mental disorders [37]. Nevertheless, one might consider using specific measures of mental capacity [38] instead of, or in conjunction with FCE tests, depending on the job requirements and the patient's impairment.
- The FCE tests in this study were used as part of a comprehensive assessment for WAD; the test results were discussed with the patient, and in some cases with the case manager. The case manager interacted with the referring occupational physician, the GP and the employer. Findings from qualitative studies suggest that consensus between the rater (clinician), patient (worker) and the employer or supervisor about the patient's WC [39,40] may be an important prerequisite for successful RTW.

IMPLICATIONS OF THE FINDINGS

Implications for the FCE rater[1]

Reliability of FCE tests

In comparison with the reliability of other physical examination procedures used in patients with neck pain [7,41]. FCE tests have shown adequate test-retest reliability in healthy and clinical populations [42-44]. Reliability values for the FCE tests used in this thesis were not

1 The term *rater* refers to the person who instructs and observes the evaluee, measures the parameters of the test and determines the level of physical effort on the basis of observational criteria. The term *rater* can be used interchangeably with *evaluator, assessor* and *tester*.

inferior to those for imaging techniques such as cervical MRI, which range from poor to moderate [2,45,46]. Nevertheless, it has to be acknowledged that on an individual level, in patients with chronic pain, the measurement error for FCE tests and self-report measures of functional ability is substantial [43,44,47-49]. For example in a FCE lifting test a change of 5kg between two time points may be due to measurement error alone [43,44]. The instrument, the patient and the rater may all be sources of error. For example, reliability may differ depending on whether a dichotomous or a four-point observational scale is used to determine physical effort [50,51]; the former may give higher intra- and inter-rater reliability estimates than the latter. Raters should also appreciate that the reliability of the observational criteria currently used in FCE tests [42,50] to determine effort during ambulation and postural tolerance tests was inadequate. Raters should be aware that their level of training may influence the reliability of a test classification [52]. FCE providers should consider providing regular training and supervision for FCE raters to improve their reliability or ensure that it is maintained at an acceptable level.

Validity of FCE tests

FCE raters must recognize that there are differences between the construct validity of FCE tests in healthy and clinical populations. In healthy populations there is a clear relationship between FCE test score and physical factors such as gender, age, aerobic capacity or muscle strength [53,54]. Conversely, in patients with chronic musculoskeletal pain functional capacity appear to be more closely related to psychosocial factors [27,29,55-58], implying that in clinical populations FCE test results reflect more than physical capacity. In patients with WAD FCE test scores are related to, but distinct from self-reported disability, mental distress and pain [59]. Thus, whilst a patient with sub-acute WAD may show high functional capacity during FCE tests despite having high self-reported disability, is less likely to be the case in a patient with chronic WAD. So, self-report measures of disability alone may be insufficient for a comprehensive evaluation.

Submaximal effort

When compensation or medico-legal outcomes are at stake submaximal effort is common [60-62]. FCE raters should be conscious that FCE tests can be biased by submaximal effort [63]. Conducting FCEs without controlling for submaximal effort could lead an inaccurate assessment of disability and thus inappropriate care, as well as unwarranted disability compensation [64]. Surprisingly the prevalence of submaximal effort is rarely reported in studies of FCE tests. The findings in this thesis confirm that FCE raters can use a lumbar motion monitor or visual observations accompanying a FCE lifting test to detect submaximal capacity in patients with chronic low back pain [65]. Additionally, the inter- and intra-tester

reliability of criteria for submaximal effort appears to be high in FCE raters [50]. Experience and additional training can increase an observer's ability to detect submaximal effort [50]. An index has been developed for reporting submaximal effort in clinical practice [66]. This submaximal effort index is moderately correlated with work status in patients with WAD [66] and so FCE raters should consider reporting submaximal effort in conjunction with other FCE results to allow patients, referrers and stakeholders to interpret the results appropriately.

FCE rater beliefs and the influence of pain behavior

The FCE rater should realize that the therapeutic alliance and rater beliefs influence the activity level of patients [67-70]. Some preliminary results with healthy undergraduate students indicated a clinically relevant increase in the weight lifted by patients tested by non-fearful raters rather than fearful raters [71]. Although these results should be replicated in clinical settings, they highlight the impact of FCE raters' beliefs on FCE test results. Pain behavior is one of the primary means by which raters infer someone's pain experience; pain behavior may also influence a rater's interpretation of a patient's ability to perform FCE tests. Patients who displayed high levels of pain behavior were judged less likely to return to work [72]; thus the rater's response to pain behavior may contribute to prolonged disability. In summary, raters should be encouraged to a) reflect critically on their beliefs about the FCE test, b) measure pain behavior when performing FCE to improve interpretation of the test results. However, at present there is no instrument validated for the measurement of pain behavior in FCE.

Cultural background of the patient

In this thesis patients with non-native mother language performed substantially worse in all FCE tests [73]. Moreover, the mother language was confirmed as independent predictor for future work status [66]. Although our findings should be replicated, these results indicate that the cultural background of the patient may play a role in FCE testing. Workers with non-native mother language are growing part of the work force in industrialized countries [74]. Workers with non-native mother language are more vulnerable and at risk for being exposed to adverse working conditions [75] and therefore they may have more difficulty in returning to work. Contradictory expectations, communication problems, mistrust in health care providers which can result in drop-out from occupational rehabilitation [76,77] may also contribute to the higher risk of N-RTW in these populations.

Whether provision of explicit information about the purpose of the FCE tests, explanations of the roles of the rater, the patient, and referrer and improving communication by using professional translators would influence FCE test results should be further studied. In summary, clinicians should be cautious when interpreting the results of studies of FCE which did not include non-native workers, as the results may not generalize to their patient population.

Table 9.1 Rationale for requesting FCE tests

Rationale	Description of test FCE objectives
1. Injury prevention	Assess whether the patient meets the requirements of the current job or not. Examples: fire fighters, emergency department personnel, workers from the urban waste disposal.
2. Management of work-related injury or illness (in context of treatment)	Assess whether the job requirements for the previous job or the adapted job are fulfilled. FCE tests can be generic or job specific.
3. Management of chronic injury and illness (usually at MMI[1])	Assess residual functional abilities in order to provide objective information for vocational training, or as a prelude to claim settlement. FCE is usually generic and includes tests to determine level and consistency of effort.

[1] MMI: maximal medical improvement: is the point at this point no further improvement is likely given the available medical and surgical treatments.' [114] [p. 22]. Table adapted from Galper E, Isernhagen S. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability. How to request, interpret, and apply functional capacity evaluations. Copyright 2009, reproduced with permission from the American Medical Association.

Comparing FCE outcomes of different worker populations

The patients tested in the research for this thesis performed substantially worse in material handling tests than 40 patients with WAD tested in a study in the Netherlands (Dutch patients lifted a mean of 12.2 kg more) [59]. The differences between the studies may be partly explained by sample variation since the Dutch study used a small sample. The difference in purposes served by the FCE tests (Table 9.1) is another possible reason for differences in results. In Switzerland FCE tests are used determine work capacity for RTW in general, whereas in the Netherlands RTW is established by the worker and his employer. Nevertheless further investigation is needed to determine whether these differences between patients with WAD were due to differences in raters, patients or the social security context. These results are consistent with a study that reported large differences between different countries in FCE outcomes in patients with chronic low back pain (CLBP) [17]; the mean capacity of patients in Canadian and Swiss samples was consistently lower than that of a Dutch sample. This association remained statistically significant after controlling for potential confounders such as age and gender. These findings emphasise the importance of exercising caution in drawing conclusions from studies in other countries and other contexts.

Comparison of FCE outcomes in patients with WAD, workers with osteoarthritis and healthy workers

It has been proposed that normative values from healthy workers should be considered as an additional screening tool for comparing workload and capacity data [78]. A comparison

of the results of FCE tests in patients tested for this thesis and healthy workers showed that in material handling tests the average score for patients with WAD was between the 5th percentile (for 'very/heavy work'; Dictionary of Occupational Titles (DOT) classification of workloads) and the 30th percentile (for 'sedentary work') of the score distribution for healthy workers [79]. Similarly, in patients with CLBP participating in a rehabilitation program [[81] and patients with early osteoarthritis [80,81], the majority of patients had FCE test scores below the lowest category of the DOT category of their healthy counterparts. Although age and gender differences may have contributed to this effect in the study of patients with CLBP [81], this was not the case in the study of patients with osteoarthritis [80,81]. Several factors besides loss of conditioning may be associated with lower FC. It was hypothesized, that patients with chronic pain stop FCE tests before maximum capacity is reached because of their experience of pain, fear of movement etc. However, self-reported pain was not associated with lower FC [81]. Another explanation is that patients may subconsciously or consciously adjust their performance to the context. Higher FC result in patients being viewed as 'recovered' and their compensation correspondingly reduced. Explanations for low FC in patients with WAD should be explored in future research. It is debatable whether the suggested normative data are appropriate to the population of interest [80]. Raters should acknowledge the advantages and the limitations of normative values in a clinical context [82] and should be encouraged to develop norms for specific clinical populations.

Implications for the patient[2]

Safety of FCE tests

The anecdotal accounts of patients, attorneys, proxies, worried health care providers and referrers indicate that FCE tests are potentially hazardous and may lead to injuries [83]. The safety of FCE tests in patients with WAD was therefore considered in the research for this thesis. In line with the results of other studies in both healthy workers and patients with CLBP [44,84,85], this research indicated that a temporary increase in pain after the FCE test followed by a return to pre-test level within a few days is common. No severe side-effects were reported by patients with WAD, nor did the increase in pain result in increased use of health care services or medication [44]. While the reasons for the temporary increase in pain are still unclear, some have termed this phenomenon 'repetition-induced summation of activity-related pain' (RISP). Early findings indicate that patients with a greater tendency to catastrophize and higher levels of fear show higher RISP during lifting tasks with constant

2 The term *patient* refers to the individual who is performing the FCE tests. The term *patient* can be used interchangeably with *evaluee, client* or *claimant*. The term *patient* was used in the studies of this thesis.

physical demands [86,87]. Delayed onset muscle soreness (DOMS) is another possible explanation for the increase in pain after FCE. DOMS usually arises after intensive physical activity associated with unfamiliar and eccentric muscular work [88]. In a study in healthy workers, 82% reported pain after FCE, and 85% reported pain that could be attributed to DOMS [89].

In conclusion, patients should be encouraged to discuss their fears and concerns with FCE raters prior testing. They may be informed by raters that FCE tests may increase pain, but that pain will usually decrease to pre-test levels within days and should not be interpreted as a sign of injury. Patients should be encouraged to report to the rater why they did not use maximal effort on a particular FCE. This information may help the rater interpret the test results more accurately.

FCE from the patient perspective

In the clinical setting used in the research for this thesis a large proportion of patients with WAD report high levels of mental distress and pain, and low functional ability [73]. Often patients have avoided physical activities for months, sometimes this avoidance has been reinforced by advice from attorneys, health care providers and proxies. Because of this patients may feel insecure about performing FCE tests if they are not told about the aim and the content of the evaluation. Patients usually perform better on FCE tests in a re-test [42-44,90]. This finding supports the clinical evidence that at least some patients are reassured by achieving certain the level of activity without severy consequences on their first FCE and therefore increase their effort on re-test, because their fear of injury is reduced. It is unclear whether patients perceive this increase in functional capacity as a 'therapeutic effect' in terms of a self-efficacy and acceptance framework [91,92]. There have been only a few studies of utility from the patient perspective [93-95]. The results of semi-structured interviews with 19 patients revealed that patients thought that FCE tests 1) altered patients' views on their ability to do physical work, 2) confirmed patients' opinions of their limitations, 3) altered the referrers view, 4) tested their physical ability, and 5) helped to determine a suitable RTW trajectory [95]. These findings suggest that FCE tests can influence patients' perspectives on abilities (self-efficacy beliefs), an effect which would be unlikely to occur if evaluation relied solely on self-report or a clinician's assessment. Self-efficacy is associated with functional capacity [27], so it is suggested that the purpose and content of the FCE should be communicated clearly to a patient to allow the patient to decide whether he or she feels able to undertake FCE tests.

Implications for referrers[3]

Determining work capacity from FCE tests

The determination of current and future work capacity is complicated; diagnosis cannot be mapped simply to functional limitations. The way in which FCE tests supplement medical records and self-report measures of fitness-for-work has been demonstrated in studies in various jurisdictions [96,97]. In a qualitative study of the utility of FCE for RTW specialists reported the following main benefits of FCE: 1) gives an overview of current physical ability of the patient; 2) verifies the consistency of verbal information provided by the patient and the patients' physical performance; 3) gives insight into the likely timescale for RTW; 4) determines the RTW trajectory [95]. Similar findings were reported in another study which asked insurance physicians and RTW specialists about the utility of FCE [98].

In spite of the limited ability of FCE tests to predict long-term RTW, referrers may expect the impossible, and some FCE providers promise the impossible i.e. that they can predict future work status of patients on the basis of FCE tests alone. There is strong evidence that lifting tests may account for 10–27% of the variance in RTW outcome at 12-month follow-term in patients with chronic musculoskeletal disorders [99-102], but in the research for this thesis FCE tests were not or were only weak predictors of short-term future work status in patients with sub-acute WAD [66]. Several other factors may also influence future work capacity e.g. the natural course of the disorder within the insurance system, cultural background, and self-reported disability. Referrers should be informed about the amount of variance in predictions of future work capacity that is accounted for by FCE tests. The main purpose of FCE tests is not to *predict* future work status but to assess *current* ability to perform work-related tasks [103,104]. The following statement underlines the limitations of FCE tests as a predictor future work status: "It is critical to understand that an instrument measuring a single dimension cannot be expected to assess a multidimensional construct. It is, therefore, by definition incorrect to suggest or to claim that the results of an FCE should be able to predict a person's work ability, or even more complex, a *successful* return to work. At best, one may expect an FCE to measure an individual's immediate functional ability to perform work-related activities" [105] [p. 106]. Referrers should discuss the aims and purposes of FCE with FCE providers and then decide whether a FCE is appropriate. The goal of a 100% accurate RTW prediction based solely on a small number of FCE tests will probably remain unattainable owing to methodological constraints, and the complexity of the interactions among the biological, psychological and social factors related to any disease or complaint [106,107]. At present, in the absence of evidence that there is a better method of assessing

3 The term *referrers* is used to refer to stakeholders who may request a FCE e.g. insurers, case managers, occupational or insurance physicians, general practitioners and other RTW experts.

patients' physical functional capability in terms of the physical demands of a given job, judicious use of FCE tests based on a thorough understanding of the specific job requirements is probably the best method of RTW trajectory [108].

Use self-report measure of functioning instead of FCE tests?

One might argue that as there is overlap between self-report measures of functioning and FCE tests, self-report measures should replace FCE tests, which are claimed to be more time consuming and can cause a temporary increase in pain. However, the overlap between self-report measures and FCE tests varies across populations. Weak relationships between FCE and self-report measures have been reported in Dutch populations of patients with chronic pain [31,32,109], but the relationship was substantially higher in Canadian and Swiss populations [73,110,111]. In a cross-sectional study for this thesis self-reported functioning measured with the SFS was moderately correlated with scores on the material handling and postural tolerance FCE tests, indicating some overlap between the two constructs [73]. However, self-reported functioning was not an independent predictor of work capacity over a 12-month period [66]. Another study produced similar preliminary findings when RTW was estimated from either FCE tests or an interview-based functional assessment based on FCE items which was administered by experienced FCE raters [112]. Although confirmation of these findings in other populations is required, they suggest – unsurprisingly, given their respective limitations – that FCE tests cannot substitute for self-report measures and *vice versa*.

In summary, it is recommended that referrers considering a FCE should focus on the following [113]: a) the characteristics of the individual who will undertake the FCE tests ('who'), b) the type of information the FCE tests are intended to provide and motivation for conducting them ('what' and 'why') and c) the context in which the FCE tests will take place ('where' and 'when') (Table 9.1). The answers to these questions should help the referrer decide whether to ask for a FCE.

RECOMMENDATIONS FOR FUTURE RESEARCH

The aim of this thesis was to contribute to the scientific knowledge about the evaluation of patients with WAD and patients with back pain using performance-based and self-report measures of functioning. Some of the pertinent gaps in FCE research were addressed [8,105,115], but other questions remain unanswered and new questions have arisen. Hence, the following recommendations for future research are proposed:

Determination of physical effort (Chapters 2 and 3)

- Carry out large validation studies using innovative study designs to develop a practical gold standard for the assessment of effort during capacity tests in clinical samples e.g. using known-groups designs and blind assessors.
- Report prevalence of submaximal effort and pain behavior alongside the FCE test results in studies of FCE.
- Validate a pain behaviour tests for FCE tests.
- Investigate the effect of including behavior tests, consistency testing or a combination of these measures [116,117] on the interpretation of FCE tests.
- Develop reliable, valid observational criteria for assessing physical effort in tests other than material handling tests e.g. overhead working, standing bend forward and ambulation.
- Determine objective-specific reliability thresholds for FCE tests for the various purposes e.g. case closure versus work disability interventions.
- Develop evidence-based guidelines for the training FCE raters.

Functional capacity tests in patients with WAD (Chapters 4, 5 and 6)

- Identify modifiable factors which influence the reliability of FCE tests.
- Use mixed methods research to characterize the influence of FCE tests on patients, raters and referrers.
- Validate FCE tests in patients with various cultural backgrounds.
- Validate FCE tests in diverse populations using job analysis and self-report measures of work-related functioning.
- Develop a cross-cultural and cross-national database of FCE test results for healthy workers, workers with pain who are at work, and various clinical populations.
- Test whether the predictive validity of FCE with respect to RTW varies across subpopulations grouped by pre-injury health, pain behavior, cultural background and the role of the work supervisor.
- Investigate whether the validity of FCE tests varies according to the particular measures to determine work capacity e.g. self-report data, insurance data, and the employer's or work supervisor's opinion.

Perceived functional ability (Chapters 7 and 8)

- Develop and evaluate a revised and shortened version of the SFS.
- Revise and validate the summary score of the SFS using job analysis, self-report measures and observation measures of functioning.

VALORIZATION

Adaptations of the WAD assessment

The studies were performed within a usual setting for the care and assessment of WAD, at the Rehaklinik Bellikon in Switzerland. Clinical concerns prompted this research project, so implementing recommendations developed from this research in clinical practice was an important goal. In the following, the clinical value and impact of the findings is discussed.

The studies reported in *Chapters 2* and *3* have led to the introduction of monitored supervision of all FCE raters by an experienced FCE rater at least once per semester. The experienced FCE rater emphasizes the importance of using the criteria for physical effort correctly, the standardized method for the interpretation of behavioral signs and the importance of the correct interpretation and consistent reporting of behavioral observations. Each FCE report now includes a clear description of the FCE test results including whether the effort on a particular FCE item was submaximal. Non-organic-somatic components were reported to be consistent independent predictors for functional capacity and therefore warranted consideration in the interpretation of FCE test results [118], so modified cervical non-organic-somatic signs for patients with neck pain were added to the WAD assessment [119].

The results reported in *Chapters 4* and *5* led to several adaptations to the original FCE test battery used in the WAD Assessment. The overhead working test was adapted so that the patient now wears hand cuffs weighing from 1 kg around both wrists. This procedure may decrease the duration of the test and reduce the number of patients who reach the ceiling (~30% of tested patients in our sample, see results in *Chapter 4*). The number of FCE tests in the assessment for WAD was reduced from 9 to 5: hand strength, lifting high, overhead working, repetitive reaching and a short walking test. These were the five tests shown to measure distinct constructs, and were valid in patients with different cultural background. The reduction in the number of tests which has to be administered produces time savings. If necessary this time can be used to conduct more job-specific tests, to assess other relevant factors such as mental capacity for work or more time can be spent on the second part of the WAD assessment: planning a brief intervention intended to enhance recovery. The FCE

tests have been retained as part of the WAD assessment, despite the fact that they do not predict RTW (Chapter 6). The primary aim of the WAD assessment was not to predict RTW, but to evaluate current functioning, and make recommendations to the patient, the referring physician or case manager, the involved health care providers and the employer. The impact of WAD assessment from the perspectives of the patient, the referrer and the employer should be explored in a future controlled study.

Tool for use in clinical practice in screening patients referred for rehabilitation

The predictive model developed in *Chapter 6* was used to develop a WC calculator. The WC calculator can be used by case managers, insurance physicians and other RTW experts to detect subjects at risk for N-RTW who may require a more integrated approach to care. Most of the information needed for the WC calculator can be drawn from a medical file; information about self-reported disability is an exception. The predictive model developed in this thesis should also be internally and externally validated.

Development of a short version of the SFS

A research project drawing on the results of *Chapters 7* and *8* to develop a revised version of the SFS was launched. This project has two parts. In the first part, a mixed methods approach was used to develop a revised version of the SFS. Data from three trials with patients with CLBP and qualitative interviews with sick-listed patients and employees working with CLBP were used to analyze SFS items. Expert opinions were also included in the assessment of the SFS. Data from all these sources were combined to select items for inclusion in the revised version of the SFS. The second part of the project will pilot the revised SFS, this will establish the test-retest reliability and construct validity of the revised SFS and the results may lead to further adaptation of the SFS.

Post-graduate education for WAD management

Unfortunately management of WAD in many patients who participated in the research for this thesis – by health care providers and by patients themselves – was often not in line with evidence based recommendations [120]. In many cases patients were not informed about the benign course of the disorder, were advised to use medication, to rest whenever pain increased and to reduce work duties. RTW was seldom recommended until pain was no longer present. In Switzerland there are currently no interdisciplinary post-graduate courses on management of WAD for health care providers and RTW experts with a strong

focus on occupational and contextual factors. In the light of this evidence and the results reported in this thesis, a post-graduate course on WAD management was introduced in November 2013. The interdisciplinary teaching team is made up of tutors from the fields of physiotherapy, psychology, medicine and case management. The aims of this course is to educate students about the mechanisms underlying the development of WAD and to determine effective strategies – particularly RTW strategies – for patients with WAD, based on an integrated approach to care which considers curative, rehabilitative and occupational medicine. An address list of course participants will help referrers to find qualified experts in the field of WAD management.

Network and knowledge center for WAD

Feedback from alumni of the post-graduate course on WAD management indicated that an easily accessible knowledge platform was required. This led to the development of the open-access online blog *NECKactive* (http://neckactive.wordpress.com). It is expected that NECKactive will improve communication among alumni of the post-graduate course on WAD management and other professions involved in WAD rehabilitation. NECKactive is intended to provide a forum for visitors and a group of authors to post the latest news on research and the best currently available evidence-based treatments for WAD. Blog participants are invited to contribute to NECKactive by sharing and discussing their experiences of managing patients with WAD.

FINAL CONCLUSIONS

This thesis determined the measurement properties of FCE tests in patients with WAD. FCE can be used to determine whether submaximal capacity when maximal capacity is asked. Clinicians can reliably determine whether submaximal effort is used during FCE tests. Four-point scales for physical effort did not achieve acceptable levels of intra-and inter-tester reliability. The safety, test-retest reliability and validity of FCE tests are acceptable when used in patients with WAD. FCE tests are related to work capacity over a short time period but do not predict work capacity over a 12-months period. Picture-based self-report measures which assess patients' perceptions of their limitations in activities involving the spine are a reliable and valid addition to FCE tests.

REFERENCES

1. Kongsted A, Sorensen JS, Andersen H, Keseler B, Jensen TS, Bendix T. Are early MRI findings correlated with long-lasting symptoms following whiplash injury? A prospective trial with 1-year follow-up. Eur Spine J. 2008;17:996-1005.

2. Anderson SE, Boesch C, Zimmermann H, Busato A, Hodler J, Bingisser R, et al. Are there cervical spine findings at MR imaging that are specific to acute symptomatic whiplash injury? A prospective controlled study with four experienced blinded readers. Radiology. 2012;262:567-75.

3. Ulbrich EJ, Anon J, Hodler J, Zimmermann H, Sturzenegger M, Anderson SE, et al. Does normalized signal intensity of cervical discs on T2 weighted MRI images change in whiplash patients? Injury. 2014;45:784-91.

4. Matsumoto M, Ichihara D, Okada E, Toyama Y, Fujiwara H, Momoshima S, et al. Modic changes of the cervical spine in patients with whiplash injury: a prospective 11-year follow-up study. Injury. 2013;44:819-24.

5. Mechanic D. The concept of illness behavior. J Chronic Dis. 1962;15:189-94.

6. OECD. Sickness, Disability and Work: Breaking the Barriers. Norway, Poland and Switzerland. OECD Publishing Paris, 2006.

7. Nordin M, Carragee EJ, Hogg-Johnson S, Weiner SS, Hurwitz EL, Peloso PM, et al. Assessment of neck pain and its associated disorders: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine (Phila Pa 1976). 2008;33:S101-22.

8. Genovese E, Galper JS. Conclusions and agenda for the future. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations. Chicago IL: American Medical Association, 2009:421-36.

9. Galper JS, Genovese E. Choosing a functional capacity evaluation evaluator. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations: American Medical Association, 2009.

10. Streiner DL, Norman GR. Health measurement scales. A practical guide to their deverlopment and use. 4th ed. Oxford: Oxford University Press, 2008.

11. de Vet HC, Terwee CB, Mokkink LB, Knol D. Measurement in Medicine: a practical guide. 1st ed. Cambridge: Cambridge University Press, 2011.

12. Sloots M. Drop-out from rehabilitation in non-native patients with chronic non-specific low back pain. PhD thesis. Amsterdam: Vrije Universiteit Amsterdam, 2010.

13. Burrus C, Ballabeni P, Deriaz O, Gobelet C, Luthi F. Predictors of nonresponse in a questionnaire-based outcome study of vocational rehabilitation patients. Arch Phys Med Rehabil. 2009;90:1499-505.

14. Scheppers E, van Dongen E, Dekker J, Geertzen J. Potential barriers to the use of health services among ethnic minorities: a review. Fam Pract. 2006;23:325-48.

15. Verbunt JA, Huijnen IP, Seelen HA. Assessment of physical activity by movement registration systems in chronic pain: methodological considerations. Clin J Pain. 2012;28:496-504.

16. Schrader H, Obelieniene D, Bovim G, Surkiene D, Mickeviciene D, Miseviciene I, et al. Natural evolution of late whiplash syndrome outside the medicolegal context. Lancet. 1996;347:1207-11.

17. Reneman MF, Kool J, Oesch P, Geertzen JH, Battie MC, Gross DP. Material handling performance of patients with chronic low back pain during functional capacity evaluation: a comparison between three countries. Disabil Rehabil. 2006;28:1143-9.

18. Sullivan MJ, Thibault P, Simmonds MJ, Milioto M, Cantin AP, Velly AM. Pain, perceived injustice and the persistence of post-traumatic stress symptoms during the course of rehabilitation for whiplash injuries. Pain. 2009;145:325-31.

19. Wenzel HG, Vasseljen O, Mykletun A, Nilsen TI. Pre-injury health-related factors in relation to self-reported whiplash: longitudinal data from the HUNT study, Norway. Eur Spine J. 2012;21:1528-35.

20. Tompa E. Measuring the burden of work disability: a review of methods, measurement issues and evidence. In Loisel P, Anema JR eds. Handbook of Work Disability. Prevention and Management. New York: Springer, 2013:43-58.

21. Wasiak R, Young AE, Roessler RT, McPherson KM, van Poppel MN, Anema JR. Measuring return to work. J Occup Rehabil. 2007;17:766-81.

22. Dasinger LK, Krause N, Deegan LJ, Brand RJ, Rudolph L. Duration of work disability after low back injury: a comparison of administrative and self-reported outcomes. Am J Ind Med. 1999;35:619-31.

23. Fadyl J, McPherson K. Return to work after injury: a review of evidence regarding expectations and injury perceptions, and their influence on outcome. J Occup Rehabil. 2008;18:362-74.

24. Rainville J, Pransky G, Indahl A, Mayer EK. The physician as disability advisor for patients with musculoskeletal complaints. Spine (Phila Pa 1976). 2005;30:2579-84.

25. Verhagen AP, Lewis M, Schellingerhout JM, Heymans MW, Dziedzic K, de Vet HC, et al. Do whiplash patients differ from other patients with non-specific neck pain regarding pain, function or prognosis? Man Ther. 2011;16:456-62.

26. Turk DC, Okifuji A. Psychological factors in chronic pain: evolution and revolution. J Consult Clin Psychol. 2002;70:678-90.

27. van Abbema R, Lakke SE, Reneman MF, van der Schans CP, van Haastert CJ, Geertzen JH, et al. Factors associated with functional capacity test results in patients with non-specific chronic low back pain: a systematic review. J Occup Rehabil. 2011;21:455-73.

28. Isernhagen SJ. Functional capacity evaluation: rational, procedure, utility of the kinesiophysical approach. J Occup Rehabil. 1992;2:157-68.

29. Lakke SE, Wittink H, Geertzen JH, van der Schans CP, Reneman MF. Factors that affect functional capacity in patients with musculoskeletal pain: a Delphi study among scientists, clinicians, and patients. Arch Phys Med Rehabil. 2012;93:446-57.

30. Soer R, van der Schans CP, Groothoff JW, Geertzen JH, Reneman MF. Towards consensus in operational definitions in functional capacity evaluation: a Delphi Survey. J Occup Rehabil. 2008;18:389-400.

31. Brouwer S, Dijkstra PU, Stewart RE, Goeken LN, Groothoff JW, Geertzen JH. Comparing self-report, clinical examination and functional testing in the assessment of work-related limitations in patients with chronic low back pain. Disabil Rehabil. 2005;27:999-1005.

32. Reneman MF, Jorritsma W, Schellekens JM, Goeken LN. Concurrent validity of questionnaire and performance-based disability measurements in patients with chronic nonspecific low back pain. J Occup Rehabil. 2002;12:119-29.

33. Balogh I, Orbaek P, Ohlsson K, Nordander C, Unge J, Winkel J, et al. Self-assessed and directly measured occupational physical activities--influence of musculoskeletal complaints, age and gender. Appl Ergon. 2004;35:49-56.

34. Amick BC, 3rd, Lerner D, Rogers WH, Rooney T, Katz JN. A review of health-related work outcome measures and their uses, and recommended measures. Spine (Phila Pa 1976). 2000;25:3152-60.

35. Abma FI, van der Klink JJ, Bultmann U. The work role functioning questionnaire 2.0 (Dutch version): examination of its reliability, validity and responsiveness in the general working population. J Occup Rehabil. 2013;23:135-47.

36. Finger ME, Escorpizo R, Bostan C, De Bie R. Work Rehabilitation Questionnaire (WORQ): Development and Preliminary Psychometric Evidence of an ICF-Based Questionnaire for Vocational Rehabilitation. J Occup Rehabil. 2013. DOI 10.1007/s10926-013-9485-2.

37. Schandelmaier S, Fischer K, Mager R, Hoffmann-Richter U, Leibold A, Bachmann MS, et al. Evaluation of work capacity in Switzerland: a survey among psychiatrists about practice and problems. Swiss Med Wkly. 2013;143:w13890.

38. Nieuwenhuijsen K, Franche RL, van Dijk FJ. Work functioning measurement: tools for occupational mental health research. J Occup Environ Med. 2010;52:778-90.

39. Abma FI, Bultmann U, Varekamp I, van der Klink JJ. Workers with health problems: three perspectives on functioning at work. Disabil Rehabil. 2013;35:20-6.

40. Coutu MF, Baril R, Durand MJ, Cote D, Cadieux G. Clinician-patient agreement about the work disability problem of patients having persistent pain: why it matters. J Occup Rehabil. 2013;23:82-92.

41. Hanney WJ, George SZ, Kolber MJ, Young I, Salamh PA, Cleland JA. Inter-rater reliability of select physical examination procedures in patients with neck pain. Physiother Theory Pract. 2014.

42. Brouwer S, Reneman MF, Dijkstra PU, Groothoff JW, Schellekens JM, Goeken LN. Test-retest reliability of the Isernhagen Work Systems Functional Capacity Evaluation in patients with chronic low back pain. J Occup Rehabil. 2003;13:207-18.

43. Soer R, Gerrits EH, Reneman MF. Test-retest reliability of a WRULD functional capacity evaluation in healthy adults. Work. 2006;26:273-80.

44. Trippolini MA, Reneman MF, Jansen B, Dijkstra PU, Geertzen JH. Reliability and safety of functional capacity evaluation in patients with whiplash associated disorders. J Occup Rehabil. 2013;23:381-90.

45. Matsumoto M, Fujimura Y, Suzuki N, Toyama Y, Shiga H. Cervical curvature in acute whiplash injuries: prospective comparative study with asymptomatic subjects. Injury. 1998;29:775-8.

46. Stafira JS, Sonnad JR, Yuh WT, Huard DR, Acker RE, Nguyen DL, et al. Qualitative assessment of cervical spinal stenosis: observer variability on CT and MR images. AJNR Am J Neuroradiol. 2003;24:766-9.

47. Smeets RJ, Hijdra HJ, Kester AD, Hitters MW, Knottnerus JA. The usability of six physical performance tasks in a rehabilitation population with chronic low back pain. Clin Rehabil. 2006;20:989-97.

48. Borloz S, Trippolini MA, Ballabeni P, Luthi F, Deriaz O. Cross-cultural adaptation, reliability, internal consistency and validation of the Spinal Function Sort (SFS) for French- and German-speaking patients with back complaints. J Occup Rehabil. 2012;22:387-93.

49. van Ittersum MW, Bieleman HJ, Reneman MF, Oosterveld FG, Groothoff JW, van der Schans CP. Functional capacity evaluation in subjects with early osteoarthritis of hip and/or knee; is two-day testing needed? J Occup Rehabil. 2009;19:238-44.

50. Trippolini MA, Dijkstra PU, Jansen B, Oesch P, Geertzen JH, Reneman MF. Reliability of Clinician Rated Physical Effort Determination During Functional Capacity Evaluation in Patients with Chronic Musculoskeletal Pain. J Occup Rehabil. 2013. DOI 10.1007/s10926-013-9470-9.

51. Isernhagen SJ, Hart DL, Matheson LM. Reliability of independent observer judgments of level of lift effort in a kinesiophysical Functional Capacity Evaluation. Work. 1999;12:145-50.

52. Werneke MW, Deutscher D, Hart DL, Stratford P, Ladin J, Weinberg J, et al. McKenzie lumbar classification: inter-rater agreement by physical therapists with different levels of formal McKenzie postgraduate training. Spine (Phila Pa 1976). 2013;39:E182-90.

53. Lakke SE, Soer R, Geertzen JH, Wittink H, Douma RK, van der Schans CP, et al. Construct validity of functional capacity tests in healthy workers. BMC Musculoskelet Disord. 2013;14:180.

54. Schenk P, Klipstein A, Spillmann S, Stroyer J, Laubli T. The role of back muscle endurance, maximum force, balance and trunk rotation control regarding lifting capacity. Eur J Appl Physiol. 2006;96:146-56.

55. Smeets RJ, Wittink H, Hidding A, Knottnerus JA. Do patients with chronic low back pain have a lower level of aerobic fitness than healthy controls?: are pain, disability, fear of injury, working status, or level of leisure time activity associated with the difference in aerobic fitness level? Spine (Phila Pa 1976). 2006;31:90-7; discussion 8.

56. Gross DP. Are functional capacity evaluations affected by the patient's pain? Curr Pain Headache Rep. 2006;10:107-13.

57. Wittink H, Michel TH, Sukiennik A, Gascon C, Rogers W. The association of pain with aerobic fitness in patients with chronic low back pain. Arch Phys Med Rehabil. 2002;83:1467-71.

58. Novy DM, Simmonds MJ, Lee CE. Physical performance tasks: what are the underlying constructs? Arch Phys Med Rehabil. 2002;83:44-7.

59. van der Meer S, Reneman MF, Verhoeven J, van der Palen J. Relationship between self-reported disability and functional capacity in patients with Whiplash Associated Disorder. J Occup Rehabil. 2013. DOI 10.1007/s10926-013-9473-6.

60. Greve KW, Ord JS, Bianchini KJ, Curtis KL. Prevalence of Malingering in Patients With Chronic Pain Referred for Psychologic Evaluation in a Medico-Legal Context. Arch Phys Med Rehabil. 2009;90:1117-26.

61. Fishbain DA, Cutler R, Rosomoff HL, Rosomoff RS. Chronic pain disability exaggeration/malingering and submaximal effort research. Clin J Pain. 1999;15:244-74.

62. Gatchel RJ. Psychosocial factors that can influence the self-assessment of function. J Occup Rehabil. 2004;14:197-206.

63. Geisser ME, Robinson ME, Miller QL, Bade SM. Psychosocial factors and functional capacity evaluation among persons with chronic pain. J Occup Rehabil. 2003;13:259-76.

64. Sindhu BS, King PM. Assessing evaluee effort. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations: American Medical Association, 2009:195-226.

65. van der Meer S, Trippolini MA, van der Palen J, Verhoeven J, Reneman MF. Which Instruments can Detect Submaximal Physical and Functional Capacity in Patients With Chronic Nonspecific Back Pain?: A Systematic Review. Spine (Phila Pa 1976). 2013. DOI 10.1097/01.brs.0000435028.50317.33.

66. Trippolini MA, Dijkstra PU, Côté P, Scholz-Odermatt SM, Geertzen JHB, Reneman MF. Can functional capacity tests predict future work capacity in patients with whiplash-associated disorders? Arch Phys Med Rehabil. 2014;Accepted, contingent on some revisions.

67. Ferreira PH, Ferreira ML, Maher CG, Refshauge KM, Latimer J, Adams RD. The therapeutic alliance between clinicians and patients predicts outcome in chronic low back pain. Phys Ther. 2013;93:470-8.

68. Linton SJ, Vlaeyen J, Ostelo R. The back pain beliefs of health care providers: are we fear-avoidant? J Occup Rehabil. 2002;12:223-32.

69. Hall AM, Ferreira PH, Maher CG, Latimer J, Ferreira ML. The influence of the therapist-patient relationship on treatment outcome in physical rehabilitation: a systematic review. Phys Ther. 2010;90:1099-110.

70. Miciak M, Gross DP, Joyce A. A review of the psychotherapeutic 'common factors' model and its application in physical therapy: the need to consider general effects in physical therapy practice. Scand J Caring Sci. 2012;26:394-403.

71. Lakke SE, Soer R, Geertzen JHB, Beetsma A, Reneman MF, van der Schans CP. Effect of physical therapist's attitude on lifting capacity. Physical Therapy 2014;Accepted, contingent on some revisions.

72. Martel MO, Wideman TH, Sullivan MJ. Patients who display protective pain behaviors are viewed as less likable, less dependable, and less likely to return to work. Pain. 2012;153:843-9.

73. Trippolini MA, Dijkstra PU, Geertzen JHB, Reneman MF. Validation of functional capacity evaluation in patients with whiplash-associated disorder J Occup Rehabil. 2014;Accepted, contigent on some revisions.

74. McCauley LA. Immigrant workers in the United States: recent trends, vulnerable populations, and challenges for occupational health. AAOHN J. 2005;53:313-9.

75. Ronda Perez E, Benavides FG, Levecque K, Love JG, Felt E, Van Rossem R. Differences in working conditions and employment arrangements among migrant and non-migrant workers in Europe. Ethn Health. 2012;17:563-77.

76. Sloots M, Dekker JH, Bartels EA, Geertzen JH, Dekker J. Reasons for drop-out in rehabilitation treatment of native patients and non-native patients with chronic low back pain in the Netherlands: a medical file study. Eur J Phys Rehabil Med. 2010;46:505-10.

77. Sloots M, Dekker JH, Pont M, Bartels EA, Geertzen JH, Dekker J. Reasons of drop-out from rehabilitation in patients of Turkish and Moroccan origin with chronic low back pain in The Netherlands: a qualitative study. J Rehabil Med. 2010;42:566-73.

78. Soer R, Hollak N, Deijs M, van der Woude LH, Reneman MF. Matching physical work demands with functional capacity in healthy workers: Can it be more efficient? Appl Ergon. 2014;45:1116-22.

79. Soer R, van der Schans CP, Geertzen JH, Groothoff JW, Brouwer S, Dijkstra PU, et al. Normative values for a functional capacity evaluation. Arch Phys Med Rehabil. 2009;90:1785-94.

80. Bieleman HJ, van Ittersum MW, Groothoff JW, Oostveen JC, Oosterveld FG, van der Schans CP, et al. Functional capacity of people with early osteoarthritis: a comparison between subjects from the cohort hip and cohort knee (CHECK) and healthy ageing workers. Int Arch Occup Environ Health. 2010;83:913-21.

81. Soer R, de Vries HJ, Brouwer S, Groothoff JW, Geertzen JH, Reneman MF. Do workers with chronic nonspecific musculoskeletal pain, with and without sick leave, have lower functional capacity compared with healthy workers? Arch Phys Med Rehabil. 2012;93:2216-22.

82. Douma RK, Soer R, Krijnen WP, Reneman M, van der Schans CP. Reference values for isometric muscle force among workers for the Netherlands: a comparison of reference values. BMC Sports Sci Med Rehabil. 2014;6:10.

83. Kersnovske S, Gibson L, Strong J. Item validity of the physical demands from the Dictionary of Occupational Titles for functional capacity evaluation of clients with chronic back pain. Work. 2005;24:157-69.

84. Reneman MF, Kuijer W, Brouwer S, Preuper HR, Groothoff JW, Geertzen JH, et al. Symptom increase following a functional capacity evaluation in patients with chronic low back pain: an explorative study of safety. J Occup Rehabil. 2006;16:197-205.

85. Soer R, Groothoff JW, Geertzen JH, van der Schans CP, Reesink DD, Reneman MF. Pain response of healthy workers following a functional capacity evaluation and implications for clinical interpretation. J Occup Rehabil. 2008;18:290-8.

86. Sullivan MJ, Thibault P, Andrikonyte J, Butler H, Catchlove R, Lariviere C. Psychological influences on repetition-induced summation of activity-related pain in patients with chronic low back pain. Pain. 2009;141:70-8.

87. Sullivan MJ, Lariviere C, Simmonds M. Activity-related summation of pain and functional disability in patients with whiplash injuries. Pain. 2010;151:440-6.

88. Cheung K, Hume P, Maxwell L. Delayed onset muscle soreness : treatment strategies and performance factors. Sports Med. 2003;33:145-64.

89. Soer R, Geertzen JH, van der Schans CP, Groothoff JW, Reneman MF. Can muscle soreness after intensive work-related activities be predicted? Clin J Pain. 2009;25:239-43.

90. Reneman MF, Dijkstra PU, Westmaas M, Goeken LN. Test-retest reliability of lifting and carrying in a 2-day functional capacity evaluation. J Occup Rehabil. 2002;12:269-75.

91. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev. 1977;84:191-215.

92. Wicksell RK, Ahlqvist J, Bring A, Melin L, Olsson GL. Can exposure and acceptance strategies improve functioning and life satisfaction in people with chronic pain and whiplash-associated disorders (WAD)? A randomized controlled trial. Cogn Behav Ther. 2008;37:169-82.

93. Gibson L, Strong J. Safety issues in functional capacity evaluation: findings from a trial of a new approach for evaluating clients with chronic back pain. J Occup Rehabil. 2005;15:237-51.

94. Strong S, Baptiste S, Clarke J, Cole D, Costa M. Use of functional capacity evaluations in workplaces and the compensation system: a report on workers' and report users' perceptions. Work. 2004;23:67-77.

95. Pas LW, Kuijer PP, Wind H, Sluiter JK, Groothoff JW, Brouwer S, et al. Clients' and RTW experts' view on the utility of FCE for the assessment of physical work ability, prognosis for work participation and advice on return to work. Int Arch Occup Environ Health. 2013;87:331-8.

96. Wind H, Gouttebarge V, Kuijer PP, Sluiter JK, Frings-Dresen MH. Complementary value of functional capacity evaluation for physicians in assessing the physical work ability of workers with musculoskeletal disorders. Int Arch Occup Environ Health. 2009;82:435-43.

97. Oesch PR, Kool JP, Bachmann S, Devereux J. The influence of a Functional Capacity Evaluation on fitness for work certificates in patients with non-specific chronic low back pain. Work. 2006;26:259-71.

98. Wind H, Gouttebarge V, Kuijer PP, Sluiter JK, Frings-Dresen MH. The utility of functional capacity evaluation: the opinion of physicians and other experts in the field of return to work and disability claims. Int Arch Occup Environ Health. 2006;79:528-34.

99. Kuijer PP, Gouttebarge V, Wind H, van Duivenbooden C, Sluiter JK, Frings-Dresen MH. Prognostic value of self-reported work ability and performance-based lifting tests for sustainable return to work among construction workers. Scand J Work Environ Health. 2012;38:600-3.

100. Streibelt M, Blume C, Thren K, Reneman MF, Mueller-Fahrnow W. Value of functional capacity evaluation information in a clinical setting for predicting return to work. Arch Phys Med Rehabil. 2009;90:429-34.

101. Gross DP, Battie MC. Functional capacity evaluation performance does not predict sustained return to work in claimants with chronic back pain. J Occup Rehabil. 2005;15:285-94.

102. Branton EN, Arnold KM, Appelt SR, Hodges MM, Battie MC, Gross DP. A short-form functional capacity evaluation predicts time to recovery but not sustained return-to-work. J Occup Rehabil. 2010;20:387-93.

103. Oliveri M, Jansen T, Oesch P, Kool J. The prognostic value of functional capacity evaluation in patients with chronic low back pain: part 1: timely return to work. And part 2: sustained recovery. Spine (Phila Pa 1976). 2005;30:1232-3; author reply 3-4.

104. Isernhagen SJ. Introduction to Functional Capacity Evaluation. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations. Chicago IL: American Medical Association, 2009:1-18.

105. Reneman M, Wittink H, Gross DP. The scientific status of functional capacity evaluation. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations: American Medical Association, 2009:393-420.

106. Reneman MF, Soer R. Was predictive validity of a job-specific FCE established? J Occup Environ Med. 2010;52:1145; author reply -6.

107. Cheng AS, Cheng SW. The predictive validity of job-specific functional capacity evaluation on the employment status of patients with nonspecific low back pain. J Occup Environ Med. 2010;52:719-24.

108. Schonstein E, Mahmud N, Kenny DT. Postoffer functional testing for injury prevention: methodological and practical considerations. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations. Chicago IL: American Medical Association, 2009:259-72.

109. Schiphorst Preuper HR, Reneman MF, Boonstra AM, Dijkstra PU, Versteegen GJ, Geertzen JH, et al. Relationship between psychological factors and performance-based and self-reported disability in chronic low back pain. Eur Spine J. 2008;17:1448-56.

110. Gross DP, Battie MC. Construct validity of a kinesiophysical functional capacity evaluation administered within a worker's compensation environment. J Occup Rehabil. 2003;13:287-95.

111. Oesch PR, Hilfiker R, Kool JP, Bachmann S, Hagen KB. Perceived functional ability assessed with the spinal function sort: is it valid for European rehabilitation settings in patients with non-specific non-acute low back pain? Eur Spine J. 2010;19:1527-33.

112. Gross DP, Asante AK, Miciak M, Battie MC, Carroll LJ, Sun A, et al. A Cluster Randomized Clinical Trial Comparing Functional Capacity Evaluation and Functional Interviewing as Components of Occupational Rehabilitation Programs. J Occup Rehabil. 2014. DOI 10.1007/s10926-013-9491-4.

113. Genovese E, Isernhagen SJ. Approach to requesting a functional evaluation. In Genovese E, Galper JS eds. Guide to the evaluation of functional ability: how to request, interpret, and apply functional capacity evaluations. Chicago IL: American Medical Association, 2009:19-40.

114. Rondinelli R, Genovese E, Kaths RT. Guides to the Evaluation of Permanent Impairment. 6th ed. Chicago IL: American Medical Association, 2008.

115. Reneman MF, Soer R, Gross DP. Developing research on performance-based functional work assessment: report on the first international functional capacity evaluation research meeting. J Occup Rehabil. 2013;23:513-5.

116. Oliveri M, Oesch P, Jansen B. Work oriented assessment in locomotor disorders from the Swiss perspective. Proceedings of the 16th European Congress of Physical and Rehabilitation Medicine in Brugge: Edizioni Minerva Medica. Torino, 2008:5-8.

117. Bianchini KJ, Greve KW, Glynn G. On the diagnosis of malingered pain-related disability: lessons from cognitive malingering research. Spine J. 2005;5:404-17.

118. Oesch P, Meyer K, Jansen B, Mowinckel P, Bachmann S, Hagen KB. What is the role of "nonorganic somatic components" in functional capacity evaluations in patients with chronic nonspecific low back pain undergoing fitness for work evaluation? Spine (Phila Pa 1976). 2011;37:E243-50.

119. Jorritsma W, Dijkstra PU, de Vries GE, Geertzen JH, Reneman MF. Physical Dysfunction and Nonorganic Signs in Patients With Chronic Neck Pain: Explorative Study Into Interobserver Reliability and Construct Validity. J Orthop Sports Phys Ther. 2014. DOI 10.2519/jospt.2014.4715.

120. Guzman J, Haldeman S, Carroll LJ, Carragee EJ, Hurwitz EL, Peloso P, et al. Clinical practice implications of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders: from concepts and findings to recommendations. Spine (Phila Pa 1976). 2008;33:S199-213.

English summary
Nederlandse samenvatting
Deutsche Zusammenfassung

ENGLISH SUMMARY

The majority of people who are exposed to a neck-injury recover quickly and return to work within days. A syndrome related to the neck-injury is termed Whiplash-Associated Disorders (WAD). WAD is classified by five grades from 0 (no symptoms) to IV (fracture or dislocation). The individuals described in this thesis have WAD grade I or II, which refers to neck complaints and musculoskeletal signs such as reduced range of motion and point tenderness. WAD grades I and II represent more than 90% of patients with WAD. Some individuals report substantial levels of disability due to WAD for a long period of time, which can lead to considerable burden to the individual and society due high costs of health care and work loss. Similar to many other unspecific musculoskeletal disorders, WAD types I and II cannot be diagnosed with current diagnostic methods such as high resolution MRI or biomarkers. Therefore, clinicians must rely on self-reported measures, clinical assessments e.g. based on medical history and physical examination, functional testing such as Functional Capacity Evaluation (FCE), or by a combination of these methods to determine the severity of the disorder and the level of (dis)ability. FCE tests are performance based tests to measure the capacity of activities in a standardized environment. These tests are used to make recommendations for participation in work while considering the person's body functions and structures, environmental factors, personal factors and health status. FCE tests may enable patients, clinicians and return to work experts to identify potential modifiable factors and enhance its management or if needed, determine work capacity. The interpretation of FCE tests results may have important consequences for the patient. Another way of measuring functioning is the use of self-reported measures. The Spinal Function Sort (SFS) is a picture-based questionnaire, which claims to measure perceived ability to perform activities of work and daily living. Both measures, the FCE and the SFS-questionnaire have not been evaluated in patients with WAD. Major scientific gaps in the evaluation of the measurement properties of functional testing in patients with WAD have been described by the "Task Force on Neck Pain and its Associated Disorders". Summarizing, rigorous evaluation of the measurement properties of FCE tests and self-reported questionnaires on functioning in patients with WAD is warranted. Hence, the first aim of this thesis was to determine the measurement properties of FCE tests in patients with WAD. The second aim was to evaluate the measurement properties of the picture-based questionnaire SFS in patients with WAD and with back pain. In Chapters 2 to 8 of this thesis the different studies are described.

The aim of the study presented in **Chapter 2** was to identify the ability of instruments designed to detect submaximal physical or functional capacity when maximal capacity is requested in patients with non-specific chronic musculoskeletal pain. A systematic review

was performed. The literature search was performed including the following databases: Web of Knowledge (including PubMed and Cinahl) Scopus and Cochrane. Two reviewers independently selected the articles based on the title and abstract according to the selection criteria. Studies were included when they contained original data and when they objectified submaximal physical or functional capacity when maximal physical or functional capacity was requested. Two authors independently extracted data and rated the quality of the articles. The included studies were scored according to the subscales "criterion validity" and "hypothesis testing" of the COSMIN checklist. A Best Evidence Synthesis was performed. From 7 studies included, 5 used a reference standard for submaximal capacity. Three studies were of good methodological quality and validly detected submaximal capacity with specificity rates between 75% and 100%. In conclusions, there is evidence that submaximal capacity can be detected in patients with chronic low back pain with visual observations accompanying a FCE lifting test or a lumbar motion monitor.

The aim of the study described in **Chapter 3** is to analyze the reliability of physical effort determination using observational criteria during FCE tests. In this study 21 raters assessed physical effort in 18 video-recorded FCE tests independently on two occasions, 10 months apart. Physical effort was rated on a categorical four-point Physical Effort Determination Scale (P_{ED}) based on the Isernhagen criteria, and a dichotomous Submaximal Effort Determination scale (S_{ED}). Cohen's Kappa, squared weighted Kappa and % agreement were calculated. Kappa values for intra-rater reliability of P_{ED} and S_{ED} for all FCE tests were moderate and substantial respectively. Kappa values for inter-rater reliability of P_{ED} for all FCE tests increased from moderate in the first to substantial in the second session for P_{ED} and was substantial in both sessions for S_{ED}. The inter-rater reliability of P_{ED} ranged from poor to almost perfect agreement between the single FCE tests. Acceptable reliability scores ($\kappa > 0.60$, agreement $\geq 80\%$) for each FCE test were observed in 38% of scores for P_{ED} and 67% for S_{ED}. On average material handling tests had a higher reliability than postural tolerance and ambulatory tests. The study concluded that dichotomous ratings of submaximal effort are more reliable than categorical criteria to determine physical effort in FCE tests. Regular education and training may improve the reliability of observational criteria for effort determination.

The study described in **Chapter 4** aimed to evaluate the test-retest reliability and safety of FCE tests for patients with WAD. Thirty-two participants (11 females and 21 males; mean age 39.6 years) with WAD grade I or II were included. The FCE consisted of 12 tests, including material handling, hand grip strength, repetitive arm movements, static arm activities, walking speed, and a 3 min step test. Overall the FCE duration was 60 minutes. The test-retest interval was seven days. Interclass correlations (model 1) (ICCs) and limits of agreement (LoA) were

calculated. Safety was assessed by a Pain Response Questionnaire, observation criteria and heart rate monitoring. ICCs ranged between moderate (3 min step test) and excellent (short two-handed carry). LoA relative to mean performance ranged between 15% (50 m walking test) and 57% (lifting waist to overhead). Pain reactions after WAD FCE decreased within days. Observations and heart rate measurements fell within the safety criteria. The test-retest reliability of the WAD FCE was moderate in two tests, good in 5 tests and excellent in 5 tests. Safety-criteria were fulfilled. Interpretation at the patient level should be performed with care because LoA were substantial.

The objective of the study, presented in **Chapter 5** was to analyse the validity of FCE in patients with WAD with different mother languages (i.e. cultural backgrounds) within a workers' compensation setting. In a cross-sectional study, 314 participants (42% females, mean age 36.7 years) with WAD (Grade I and II) were referred for an interdisciplinary assessment including FCE tests. Four FCE tests (hand grip strength, lifting waist to overhead, overhead working, and repetitive reaching) and a number of concurrent variables such as self-reported pain, capacity, disability, and psychological distress were measured. To test construct validity, hypotheses concerning FCE and gender, and FCE and other self-reported measures on pain, perceived functional ability, disability and mental distress were formulated a priori and tested with correlations. Men had significantly greater hand grip strength (+17.5 kg) and lifted more weight (+3.7 kg) than women. Three out of four gender-related hypotheses were not rejected. Correlation was low between FCE and pain; was moderate between FCE and perceived functional ability; was low between FCE and disability; was low between FCE and anxiety and depression, respectively. Hypotheses regarding FCE and other self-reported measures were not rejected in 16 of 20 hypotheses. FCE test results between groups of different mother language differed significantly in six out of eight FCE tests. In 18 out of 20 analyses, correlations between FCE and self-reported measures did not differ between groups of different mother language. The conclusions of this study were: the validity of FCE is good for testing functional capacity in patients with WAD with different cultural backgrounds and in a workers' compensation setting. Additional validation studies in other settings are needed for verification.

The aim of **Chapter 6** was to determine whether FCE tests can be used to predict future work capacity of patients with WAD. A prospective cohort study was performed in an outpatient work rehabilitation center. Sick listed workers with WAD 6 to 12 weeks after injury were included in the study. These patients performed 8 work-related FCE tests. The outcome work capacity (WC; 0–100%) was measured at baseline and 1, 3, 6, and 12 months after testing. Correlation coefficients between FCE tests and WC were calculated. A linear mixed model analysis was used to assess the association between FCE and future WC. In total 267 patients

with WAD grade I or II participated in the study. Mean WC increased over time from 20.8% at baseline to 83.2% at 12 months follow-up, respectively. Correlation coefficients between FCE tests and WC ranged between little if any for lifting low at 12 months follow-up to weak for walking speed at 3 months. Strength of correlations decreased over time. FCE tests did not predict WC at follow-up. The predictors of WC were ln (time) (β=23.74), mother language (β=5.49), work capacity at baseline (β=1.01), and self-reported disability (β=-0.20). Two interaction terms ln (time) x WC (β=-0.19), and ln (time) x self-reported disability (β=-0.21) were significant predictors of WC. In conclusion, FCE tests performed within 3 months after WAD injury are associated to WC at baseline, but do not predict future WC, whereas time course, mother language, WC at baseline, and self-reported disability do predict future WC. Additionally, interaction between time course WC at baseline and self-reported disability respectively predicted future WC.

The aim of the study in **Chapter 7** was to thoroughly analyze the measurement properties of the SFS in patients with sub-acute WAD grade I and II. Three-hundred-two patients with WAD were recruited from an outpatient work rehabilitation center. Internal consistency was assessed by Cronbach's α. Construct validity was tested based on 8 a priori hypotheses. Structural validity was measured with principal component analysis (PCA). Test-retest reliability and agreement was evaluated in a sub-sample (n=32) using interclass correlation coefficient (ICC) and limits of agreement (LoA). The predictive validity of SFS for future work status at 1, 3, 6 and 12 months follow-up was determined by area under the curve (AUC) of receiver operating characteristics. Non-return to work (N-RTW) was defined with two cut-off points: WC <50% and <100%. Proportion of patients N-RTW decreased from 50% to 14% between 1 and 12 months follow-up. Cronbach's α was >0.90, PCA revealed evidence for unidimensionality. ICC was 0.86 (95% CI: 0.71; 0.93), LoA was ±33 points. Seven out of 8 hypotheses for construct validity were not rejected. AUC reduced with a longer follow-up from 0.71 for 1 month to 0.68 at 12 months, for cut-off point <50%. For cut-off point <100% these values were 0.71 and 0.59. In patients with sub-acute WAD test-retest reliability, internal consistency, construct- and structural validity of the SFS were adequate. LoA were substantial. The validity to predict N-RTW was sufficient on a short term (1 month) but poor on a long term.

In **Chapter 8**, the aim was to translate, adapt and validate the French (SFS-F) and German (SFS-G) versions of the SFS. Three-hundred-forty-four patients, experiencing various back complaints, were recruited in a French (n=87) and a German-speaking (n=257) centre. Construct validity was tested via correlations with SF-36 physical and mental scales, Visual Analogue Scale (VAS) for pain and Hospital Anxiety and Depression Scales (HADS). Scale homogeneities were assessed by Cronbach's α. Test-retest reliability was assessed on 65

additional patients using intraclass correlation coefficient (ICC) and limits of agreement (LoA). For the SFS-F and SFS-G, respectively, α and ICC were >0.90 for both language versions. Correlations with physical functioning were on average moderate for both; with physical summary were moderate for both; with pain were weak for the SFS-F and moderate for the SFS-G; with mental health and mental summary were little if any for both; with depression and anxiety were weak for both versions. ICC was >0.90 for SFS-F and SFS-G. LoA were -11.5 and 12.1 for the SFS-F, and -27.7 and 30.2 points for the SFS-G. Convergent validity was good with SF-36 physical scales, moderate with VAS pain. Divergent validity was low with SF-36 mental scales in both translated versions and with HADS for the SFS-F (moderate in SFS-G). A substantial difference between LoA for the SFS-F and SFS-G was detected. Both versions seem to be valid and reliable for evaluating perceived functional capacity in patients with back complaints.

In **Chapter 9** the main findings of this thesis are discussed, integrated and reflected on. Strengths and weaknesses and other methodological considerations are discussed. Implications of the findings for the patient, the FCE rater and the referrer for FCE are described. Recommendations for future research directions are made. Adapted clinical procedures based on findings of this thesis and suggestions for further implementations are discussed in the paragraph on "valorization". Ultimately, the final conclusions are reported. This thesis provides a thorough evaluation of measurement properties of FCE tests and the SFS.

There is evidence that FCE tests can be used to determine whether a patient has performed sub-maximally or not. While the intra- and inter tester reliability of raters using observational criteria to determine the level of physical effort during FCE tests is acceptable for material handling test, this is not the case for postural tolerance and ambulation tests. FCE tests have proved acceptable test-retest reliability and construct validity in patients with WAD grades I and II with different mother languages. On an individual level, clinicians should be aware of the substantial measurement error of both the FCE tests and the SFS. FCE tests and all other modifiable variables did not predict future work capacity in patients with sub-acute WAD, whereas course of time, baseline work capacity, mother language and self-reported disability did. This challenges specific rehabilitative interventions and underlines the difficulties of influencing RTW rates. The French and German versions of the SFS appear reliable and valid for patients with back pain and patients with WAD grades I and II. Future WC of sub-acute patients with WAD grades I and II could not be predicted with the SFS, which is in contrast with findings in patients with chronic low back pain. Proposals to advance the SFS were made e.g. by halving the number of items, while the measurement properties of the SFS were not expected to alter.

The findings of this thesis may inspire clinicians and researchers to replicate these studies, and further develop the measures analyzed in this thesis. It is hoped the results of the thesis emphasize the influence of different settings or patients with different cultural background on the properties of functional measures. Hopefully, these findings will improve the way patients, health care providers and referrers interpret and implement the results of FCE tests and the SFS, and ultimately optimize the process of RTW.

NEDERLANDSE SAMENVATTING

De meeste mensen die nekletsel hebben opgelopen, herstellen snel en zijn binnen enkele dagen weer terug op hun werk. Een syndroom dat met netletsel samenhangt, is Whiplash-Associated Disorder (WAD). WAD kent een indeling in vijf graden, namelijk van 0 (geen symptomen) tot IV (fractuur of dislocatie). Alle patiënten in dit proefschrift hebben WAD-graad I of II, wat inhoudt dat ze nekklachten hebben en klachten van het bewegingsapparaat zoals bewegingsbeperkingen en drukpijn. Meer dan negentig procent van alle patiënten met WAD heeft WAD-graad I of II. Sommige patiënten geven aan dat ze als gevolg van WAD gedurende lange tijd aanzienlijke beperkingen ondervinden. Dat kan een grote belasting vormen voor de patiënt en samenleving door de hoge kosten van de gezondheidszorg en het arbeidsverzuim. Net als veel andere aspecifieke aandoeningen van het bewegingsapparaat kunnen WAD-graad I en II niet worden vastgesteld met de bestaande diagnostische methoden, zoals hogeresolutie-MRI of biomarkers. Om de ernst van deze aandoening en de mate van beperking te kunnen bepalen, moeten artsen daarom vertrouwen op zelfrapportagemethoden, klinische beoordelingen (bijv. op basis van de medische voorgeschiedenis en lichamelijk onderzoek), functionele tests zoals een Functionele Capaciteit Evaluatie (FCE), of een combinatie van deze methoden. FCE-tests zijn mede gebaseerd op fysieke prestaties en meten het vermogen van iemand om bepaalde activiteiten uit te voeren in een gestandaardiseerde omgeving. Dergelijke tests worden gebruikt om aanbevelingen te doen over de mate waarin iemand kan deelnemen aan werk, waarbij rekening wordt gehouden met zijn lichaamsfuncties en -structuren, omgevingsfactoren, persoonlijke factoren en gezondheidstoestand. FCE-tests kunnen patiënten, artsen en deskundigen op het gebied van arbeidsreïntegratie helpen factoren te achterhalen die eventueel zijn bij te sturen, zodat de behandeling van de patiënt kan worden verbeterd. Indien nodig, kunnen FCE-tests ook de arbeidscapaciteit helpen vaststellen. De interpretatie van de resultaten van FCE-tests kan belangrijke gevolgen hebben voor de patiënt. Een andere manier om het functioneren te bepalen, is met zelfrapportagemethoden. De Spinal Function Sort (SFS) is een vragenlijst op basis van afbeeldingen, die is ontworpen om te meten in hoeverre iemand denkt bepaalde handelingen te kunnen verrichten op het werk en in het dagelijkse leven. Beide instrumenten, de FCE en de SFS, zijn nog niet geëvalueerd voor patiënten met WAD. De Task Force on Neck Pain and its Associated Disorders heeft gerapporteerd dat de evaluatie van de meeteigenschappen van functionele tests bij WAD-patiënten belangrijke wetenschappelijke hiaten vertoont. Kortom, er zijn gegronde redenen om de meeteigenschappen van FCE-tests en zelfrapportagevragenlijsten voor functioneren bij patiënten met WAD zorgvuldig te evalueren. Het eerste doel van dit proefschrift was daarom het bepalen van de meeteigenschappen van FCE-tests bij patiënten met WAD. Het tweede doel was het evalueren van de meeteigenschappen van een op afbeeldingen gebaseerde

vragenlijst, de SFS, bij patiënten met WAD en patiënten met rugpijn. De verschillende studies zijn in hoofdstuk 2 tot en met 8 van dit proefschrift beschreven.

Het doel van het onderzoek in **hoofdstuk 2** was om de bruikbaarheid te bepalen van instrumenten die ontworpen zijn om de submaximale fysieke of functionele capaciteit te meten terwijl een maximale capaciteit wordt gevraagd van patiënten met aspecifieke chronische pijn aan het bewegingsapparaat. Hiervoor is een systematisch review uitgevoerd. Voor het literatuuronderzoek zijn de volgende databases gebruikt: Web of Knowledge (waaronder PubMed en Cinahl), Scopus en Cochrane. Twee onderzoekers selecteerden onafhankelijk van elkaar de artikelen op basis van de titel en het abstract volgens de gestelde selectiecriteria. Studies werden opgenomen als ze oorspronkelijke gegevens bevatten en op objectieve wijze weergaven wat de submaximale fysieke of functionele capaciteit was terwijl om een maximale fysieke of functionele capaciteit werd gevraagd. Twee auteurs verzamelden onafhankelijk van elkaar gegevens en beoordeelden de kwaliteit van de artikelen. De opgenomen studies kregen scores toegekend volgens de subschalen 'criteriumvaliditeit' en 'hypothesetoetsing' van de COSMIN-checklist, en er werd een best-evidencesynthese uitgevoerd. Van de zeven studies die werden opgenomen, gebruikten er vijf een referentiestandaard voor de submaximale capaciteit. Drie studies hadden een goede methodologische kwaliteit en bepaalden de submaximale capaciteit op valide wijze met een specificiteit van 75 tot 100%. Dit onderzoek wees uit dat de submaximale capaciteit bij patiënten met chronische lage rugpijn kan worden vastgesteld door middel van visuele observatie in combinatie met een FCE-tiltest of een bewegingsmonitor voor de onderrug.

Het onderzoek in **hoofdstuk 3** had als doel het analyseren van de betrouwbaarheid van inspanningsbeoordelingen die tijdens FCE-tests gedaan worden met behulp van observatiecriteria. In dit onderzoek keken 21 beoordelaars bij twee gelegenheden met een tussenperiode van tien maanden onafhankelijk van elkaar naar achttien gefilmde FCE-tests. Daarbij beoordeelden zij de fysieke inspanning volgens de Physical Effort Determination-schaal (P_{ED}), een categorische vierpuntsschaal op basis van de criteria van Isernhagen, en volgens de dichotome Submaximal Effort Determination-schaal (S_{ED}). Cohens kappa, de *squared weighted kappa* en het percentage overeenstemming werden berekend. De kappawaarden voor de interbeoordelaarsbetrouwbaarheid van P_{ED} waren voor alle FCE-tests gestegen van redelijk in de eerste, tot goed in de tweede sessie voor P_{ED}, en waren goed in beide sessies voor S_{ED}. De interbeoordelaarsbetrouwbaarheid van P_{ED} liet een overeenstemming tussen de afzonderlijke FCE-tests zien die varieerde van slecht tot zeer goed. Acceptabele betrouwbaarheidsscores ($\kappa > 0{,}60$; overeenstemming ≥80%) voor elke FCE-test werden gevonden in 38% van de scores voor P_{ED} en in 67% voor S_{ED}. Bij de til- en draagtests werd gemiddeld een hogere betrouwbaarheid gevonden dan bij de houdingstolerantietests

en looptests. De conclusie van dit onderzoek was dat dichotome beoordelingen van submaximale inspanning betrouwbaarder zijn om de inspanning tijdens FCE-tests te bepalen dan beoordelingen op basis van categorische criteria. De betrouwbaarheid van de observatiecriteria voor inspanningsbeoordelingen zou door regelmatige scholing en training verbeterd kunnen worden.

Het onderzoek in **hoofdstuk 4** had als doel de test-hertestbetrouwbaarheid en de veiligheid van FCE-tests te evalueren voor patiënten met WAD. Tweeëndertig deelnemers (11 vrouwen en 21 mannen; gemiddelde leeftijd 39,6 jaar) met WAD-graad I of II deden aan het onderzoek mee. De FCE bestond uit twaalf tests, waaronder tests voor tillen en dragen, handknijpkracht, repetitieve armbewegingen, statische armtaken en loopsnelheid en een 3-minuten steptest. De FCE duurde in totaal 60 minuten. Tussen de eerste en tweede testsessie lag een periode van zeven dagen. De interclass-correlatie (model 1) (ICC) en de grenzen van overeenstemming (*limits of agreement*; LoA) werden berekend, en de veiligheid werd beoordeeld met een Pain Response Questionnaire, observatiecriteria en hartslagmetingen. De ICC's varieerden van redelijk (3-minuten steptest) tot zeer goed (dragen kort). De LoA in verhouding tot de gemiddelde prestatie varieerden van 15% (50-meterlooptest) tot 57% (tillen hoog). De pijnrespons na de FCE-tests nam binnen enkele dagen af. De observaties en hartslagmetingen lagen binnen de veiligheidscriteria. De test-hertestbetrouwbaarheid van de FCE-tests was redelijk in twee tests, goed in vijf tests en zeer goed in vijf tests. Er werd aan de veiligheidscriteria voldaan. De interpretatie op het niveau van individuele patiënten moet echter met voorzichtigheid worden gedaan, omdat de LoA aanzienlijk waren.

Het doel van het onderzoek in **hoofdstuk 5** was het analyseren van de validiteit van FCE bij WAD-patiënten met verschillende moedertalen (oftewel culturele achtergronden), binnen een context van ongevallenverzekeringen voor werknemers. In een cross-sectioneel onderzoek werden 314 deelnemers (42% vrouwen, gemiddelde leeftijd 36,7 jaar) met WAD-graad I of II doorverwezen voor een interdisciplinair onderzoek dat ook FCE-tests behelsde. Er werden vier FCE-tests uitgevoerd: handknijpkracht, tillen hoog, bovenhands werk en repetitief reiken. Ook werden diverse bijkomende variabelen gemeten, zoals zelfgerapporteerde pijn, capaciteit, beperkingen en psychische stress. Voor het testen van de constructvaliditeit werden a-priorihypothesen opgesteld voor FCE en geslacht, en FCE en de overige zelfgerapporteerde uitkomstmaten voor pijn, het ervaren functionele vermogen, beperkingen en psychische stress. Deze hypothesen werden getoetst met correlaties. Mannen hadden een significant grotere handknijpkracht (17,5 kg meer) en tilden meer gewicht (3,7 kg meer) dan vrouwen. Drie van de vier hypothesen voor geslacht werden niet verworpen. De correlatie was gering tussen FCE en pijn; redelijk tussen FCE en het ervaren functionele vermogen; gering tussen FCE en beperkingen; en gering tussen FCE en respectievelijk angst en depressie. Van de

twintig hypothesen voor FCE en de overige zelfgerapporteerde uitkomstmaten werden er zestien niet verworpen. De FCE-testresultaten tussen de groepen met verschillende moedertalen waren in zes van de acht FCE-tests significant verschillend. In achttien van de twintig analyses verschilden de correlaties tussen FCE en zelfgerapporteerde uitkomstmaten niet tussen de groepen met verschillende moedertalen. De conclusies uit dit onderzoek luidden als volgt: de validiteit van FCE is goed voor het testen van de functionele capaciteit bij WAD-patiënten met verschillende culturele achtergronden en binnen een context van ongevallenverzekeringen voor werknemers. Meer validatiestudies binnen andere contexten zijn nodig om deze bevindingen te verifiëren.

Het doel van **hoofdstuk 6** was te bepalen of FCE-tests gebruikt kunnen worden om de toekomstige arbeidscapaciteit van patiënten met WAD te voorspellen. Hiervoor werd een prospectief cohortonderzoek uitgevoerd in een poliklinisch arbeidsrevalidatiecentrum. De deelnemers aan het onderzoek waren werkenden die ziek gemeld waren als gevolg van een WAD-letsel dat ze zes tot twaalf weken eerder hadden opgelopen. In totaal namen 267 patiënten met WAD-graad I of II deel aan het onderzoek. Deze patiënten voerden acht werkgerelateerde FCE-tests uit. De uitkomstmaat arbeidscapaciteit (*work capacity*, WC; 0–100%) werd gemeten bij aanvang, en 1, 3, 6 en 12 maanden na het testen. De correlatiecoëfficiënten tussen de FCE-tests en de WC werden berekend. De samenhang tussen FCE en de toekomstige WC werd bepaald met een analyse volgens het lineaire gemengde model. De gemiddelde WC steeg in de loop van de tijd van 20,8% bij aanvang, tot 83,2% bij de follow-up na twaalf maanden. De correlatiecoëfficiënten tussen de FCE-tests en de WC varieerden van 'bijna nul' voor tillen laag bij follow-up na twaalf maanden, tot 'matig' voor de loopsnelheid na 3 maanden. De sterkte van de correlaties nam in de loop van de tijd af. Geen van de FCE-tests voorspelde de WC bij follow-up. De voorspellers van de WC waren: ln(tijd) (β=23,74), moedertaal (β=5,49), arbeidscapaciteit bij aanvang (β=1,01) en zelfgerapporteerde beperkingen (β=-0,20). Ook de twee interactietermen ln(tijd) × WC (β=-0,19) en ln(tijd) × zelfgerapporteerde beperkingen (β=-0,21) waren significante voorspellers van de WC. De conclusie was dat FCE-tests die binnen 3 maanden na het WAD-letsel worden uitgevoerd, samenhang vertonen met de WC bij aanvang, maar geen voorspeller zijn van de toekomstige WC. De verstreken tijdsduur, moedertaal, WC bij aanvang en zelfgerapporteerde beperkingen voorspelden daarentegen wel de toekomstige WC. Verder waren ook de interacties tussen de verstreken tijdsduur en respectievelijk de WC bij aanvang en de zelfgerapporteerde beperkingen voorspellers van de toekomstige WC.

Het doel van het onderzoek in **hoofdstuk 7** was om de meeteigenschappen van de SFS grondig te analyseren bij patiënten met subacute WAD-graad I en II. Er werden 302 WAD-patiënten van een poliklinisch arbeidsrevalidatiecentrum opgenomen in het onderzoek.

De interne consistentie werd bepaald met Cronbachs α. De constructvaliditeit werd onderzocht met behulp van acht a-priorihypothesen. De structurele validiteit werd bepaald met de principale componenten analyse (PCA). De test-hertestbetrouwbaarheid en de overeenstemming werden beoordeeld in een subgroep (n=32) door middel van de interclass-correlatiecoëfficiënt (ICC) en de grenzen van overeenstemming (LoA). De predictieve validiteit van de SFS voor de toekomstige werksituatie bij follow-up na 1, 3, 6 en 12 maanden werd bepaald met de oppervlakte onder de curve (*area under the curve*, AUC) van de ROC-curve (*receiver operating characteristics curve*). De uitkomst 'geen terugkeer naar werk' (*non-return to work*, N-RTW) werd gedefinieerd met twee afkappunten: WC <50% en <100%. De proportie N-RTW-patiënten daalde van 50% bij follow-up na één maand, tot 14% bij follow-up na twaalf maanden. Cronbachs α was groter dan 0,90 en de PCA wees uit dat er sprake was van unidimensionaliteit. De ICC was 0,86 (95% CI: 0,71; 0,93); de LoA waren ±33 punten. Zeven van de acht hypothesen voor de constructvaliditeit werden niet verworpen. De AUC werd kleiner naarmate de follow-upduur langer werd: van 0,71 bij één maand tot 0,61 bij twaalf maanden, bij een afkappunt van <50%. Bij een afkappunt van <100% was dat respectievelijk 0,71 en 0,59. Bij patiënten met subacute WAD waren de test-hertestbetrouwbaarheid, interne consistentie, constructvaliditeit en structurele validiteit voldoende. De LoA waren echter aanzienlijk. De predictieve validiteit voor N-RTW was op de korte termijn (één maand) voldoende, maar op de lange termijn slecht.

Het doel van het onderzoek in **hoofdstuk 8** was om de SFS in het Frans (SFS-F) en Duits (SFS-G) te vertalen en te bewerken, en beide taalversies te valideren. Er werden 344 patiënten met diverse rugklachten in het onderzoek opgenomen uit een centrum waar Frans (n=87) en een waar Duits (n=257) werd gesproken. De constructvaliditeit werd onderzocht door middel van correlaties met de fysieke en mentale schalen van de SF-36-vragenlijst, de Visual Analogue Scale (VAS) voor pijn en de Hospital Anxiety and Depression Scales (HADS). De homogeniteit van deze schalen werd bepaald met Cronbachs α. In een aparte groep van 65 patiënten werd de test-hertestbetrouwbaarheid onderzocht met behulp van de intraclass-correlatiecoëfficiënt (ICC) en de grenzen van overeenstemming (LoA). Voor zowel de SFS-F als de SFS-G waren α en ICC allebei groter dan 0,90. De correlatie met het fysieke functioneren was voor beide taalversies gemiddeld genomen redelijk, en met de score voor de fysieke componenten voor beide versies ook redelijk. De correlatie met pijn was voor de SFS-F matig en voor de SFS-G redelijk. Met de mentale gezondheid en de score voor de mentale componenten was de correlatie voor beide versies bijna nul, en met depressie en angst voor beide versies matig. De ICC was voor zowel de SFS-F als de SFS-G groter dan 0,90. De LoA waren voor de SFS-F -11,5 en 12,1 punten, en voor de SFS-G -27,7 en 30,2 punten. De convergente validiteit met de fysieke schalen van de SF-36 was goed, en met de VAS voor pijn redelijk. De divergente validiteit met de mentale schalen van de SF-36 was voor beide

vertaalde versies gering, en met de HADS voor de SFS-F ook gering (voor de SFS-G redelijk). De LoA voor de SFS-F en de SFS-G bleken aanzienlijk van elkaar te verschillen. Beide taalversies leken valide en betrouwbaar te zijn voor het evalueren van de ervaren functionele capaciteit bij patiënten met rugklachten.

In **hoofdstuk 9** worden de belangrijkste bevindingen uit dit proefschrift besproken, in een bredere context geplaatst en van diverse kanten belicht. Sterke en zwakke punten en andere methodologische overwegingen worden besproken. Ook wordt ingegaan op de implicaties van de bevindingen voor patiënten, FCE-beoordelaars en -doorverwijzers, en worden er aanbevelingen gedaan voor toekomstig onderzoek. De sectie 'Valorisatie' beschrijft diverse aanpassingen van klinische procedures die op basis van de bevindingen in dit proefschrift zijn gedaan, en geeft suggesties voor verdere praktische toepassingen. Het hoofdstuk sluit af met de eindconclusies. In dit proefschrift werden de meeteigenschappen van FCE-tests en de SFS grondig geëvalueerd.

Het onderzoek toont aan dat FCE-tests kunnen worden gebruikt om te bepalen of een patiënt al dan niet submaximaal presteert. De intra- en interbeoordelaarsbetrouwbaarheid van beoordelaars die observatiecriteria gebruiken om het fysieke inspanningsniveau tijdens FCE-tests te bepalen, zijn acceptabel voor til- en draagtests, al geldt dat niet voor houdingstolerantie- en looptests. Ook is aangetoond dat FCE-tests een acceptabele test-hertestbetrouwbaarheid en constructvaliditeit hebben bij patiënten met WAD-graad I en II die verschillende moedertalen spreken. Op het niveau van individuele patiënten moeten artsen echter rekening houden met de aanzienlijke meetfout van zowel de FCE-tests als de SFS. De FCE-tests en alle overige veranderlijke variabelen waren geen voorspellers van de toekomstige arbeidscapaciteit bij patiënten met subacute WAD, terwijl de verstreken tijdsduur, arbeidscapaciteit bij aanvang, moedertaal en zelfgerapporteerde beperkingen dat wel waren. Dit zet vraagtekens bij bepaalde revalidatie-interventies en onderstreept hoe moeilijk het is om invloed uit te oefenen op werkhervatting. De Franse en Duitse versies van de SFS blijken betrouwbaar en valide te zijn voor patiënten met rugpijn en patiënten met WAD-graad I en II. De toekomstige arbeidscapaciteit van patiënten met subacute WAD-graad I en II kon niet met de SFS worden voorspeld, wat in tegenspraak is met eerdere bevindingen bij patiënten met chronische lage rugpijn. Er werden voorstellen gedaan om de SFS te verbeteren, zoals het halveren van het aantal items, wat naar verwachting geen invloed heeft op de meeteigenschappen van de SFS.

Misschien inspireren de bevindingen in dit proefschrift artsen en onderzoekers om deze studies te herhalen en de onderzochte meetinstrumenten verder te ontwikkelen. De resultaten van dit proefschrift benadrukken wellicht welke invloed verschillende settings en patiënten met verschillende culturele achtergronden hebben op de eigenschappen van

functionele meetinstrumenten. Hopelijk zijn patiënten, zorgverleners en doorverwijzers door deze bevindingen beter in staat de resultaten van FCE-tests en SFS's te interpreteren en implementeren. Naar verwachting zal dit proefschrift een bijdrage leveren aan het optimaliseren van het werkhervattingsproces.

DEUTSCHE ZUSAMMENFASSUNG

Die Mehrheit der Personen, die ein leichtes Trauma der Halswirbelsäule erleiden, erholen sich rasch und nehmen ihre berufliche Tätigkeit innerhalb weniger Tage wieder auf. Ein Syndrom, welches mit einem Halswirbelsäulen-Trauma assoziiert ist, wird „Schleudertrauma"[1], nachfolgend Halswirbelsäulen-Distorsionstrauma (HWS-Distorsionstrauma) kurz, HD genannt. HD wird in Schweregrade von Grad 0 (keine Symptome) bis Grad IV (Fraktur oder Dislokation) eingeteilt. Die Studien-Teilnehmer dieser Dissertation haben ein HD Grad I oder II, d.h. es liegen Nacken-Schmerzen mit muskuloskeletalen Zeichen vor wie zum Beispiel eine verringerte Nackenbeweglichkeit oder Druckempfindlichkeit der Muskulatur. Mehr als 90% der Patienten mit HD haben ein Schweregrad I oder II. Einige dieser Patienten entwickeln ein chronisches HD, begleitet von zahlreichen Alltagseinschränkungen und Symptomen wie zum Beispiel Schwindelgefühle, Kopfschmerzen, Ermüdbarkeit. Das chronische HD führt wegen hohen Gesundheitskosten und länger dauernden Arbeitsabsenzen zu einer beträchtlichen Belastung für die betroffenen Personen und die Gesellschaft. HD Grad I und II können, ähnlich wie zahlreiche andere unspezifische muskuloskeletalen Beschwerden, nicht mit „objektiven" Messverfahren, wie zum Beispiel der Magnetresonanztomografie diagnostiziert werden. Daher stützt man sich im klinischen Alltag auf die Anamnese, spezifische Fragebögen und klinischen Untersuchungen inkl. Bewegungs- und Funktionstests. Zum Beispiel können Tests der Evaluation der Funktionellen Leistungsfähigkeit (EFL) helfen, ein Bild des Schweregrades der HD-Problematik zu erhalten. EFL-Tests sind standardisierte-Leistungstests mit denen die Belastbarkeit für häufige physische Funktionen der Arbeit untersucht wird wie zum Beispiel Heben, Tragen, Überkopfarbeit, Leitersteigen, Handkraft und -koordination. Die Hauptvorteile der EFL liegen in einer umfassenden und systematischen Leistungsevaluation mit arbeitsbezogenen Belastungen unter Berücksichtigung des beobachteten Leistungsverhaltens des Patienten bei den Tests. Berücksichtigt werden dabei die Körperfunktion, die Umgebungsfaktoren, die persönlichen Merkmale und der Gesundheitszustand. Die Ergebnisse von EFL-Tests können Patienten, Mediziner und Spezialisten der Arbeitswiedereingliederung unterstützen, mögliche veränderbare Defizite zu identifizieren, entsprechende rehabilitative Massnahmen gezielter durchzuführen oder auch die Arbeitsfähigkeit bzw. die sogenannte Zumutbarkeit einer Arbeitstätigkeit festzulegen. Letzteres kann bedeutende Auswirkungen für die betroffene Person haben. Eine andere Methode die Leistungsfähigkeit zu messen, ist die Verwendung von spezifischen

1 Das sog. „Schleudertrauma" hat in der deutschsprachigen Literatur viele Synonyme, welche bei genauer Prüfung alle ihre Unzulänglichkeiten haben: Halswirbelsäulen-Distorsionstrauma (HWS-Distorsionstrauma, kurz HD), Beschleunigungsverletzung der HWS, HWS-Beschleunigungstrauma, HWS-Peitschenschlagverletzung, Zervikozephales Beschleunigungstrauma, HWS-Akzelerations-Dezelerations-Trauma.

Fragebögen, in denen der Patient selbst seine Leistungsfähigkeit beurteilt. Im Fragebogen PACT (Performance Assessment of Capacity Testing; Englisch: Spinal Function Sort) beurteilen Patienten die eigene Leistungsfähigkeit bei Alltags- und Berufsaktivitäten. Jede Frage wird dabei von einem Bild illustriert. Die EFL-Tests wie auch der PACT-Fragebogen sind nicht auf ihre Gütekriterien bei Patienten mit HD untersucht worden. Die "Task Force on Neck Pain and its Associated Disorders" hat hinsichtlich der Gütekriterien von Tests, welche bei Patienten mit HD verwendet werden, grössere Wissenslücken mit entsprechendem Forschungsbedarf festgestellt. Zusammenfassend lässt sich sagen, dass eine wissenschaftliche Überprüfung der Test-Gütekriterien von EFL-Tests und des PACT-Fragebogens bei Patienten mit HD notwendig ist. Folglich ist das erste Ziel dieser Dissertation die Gütekriterien der EFL-Tests bei Patienten mit HD zu überprüfen. Zweitens sollen die Gütekriterien des bildergestützten PACT-Fragebogens bei Patienten mit HD und bei Patienten mit Kreuzbeschwerden untersucht werden. In den folgenden Kapitel 2 bis 8 werden die durchgeführten Studien beschrieben.

Kapitel 2 beschreibt eine Studie die Tests identifizieren soll, die valide sind nicht-maximaler Leistungsfähigkeit von maximaler Leistungsfähigkeit von Patienten mit unspezifischen, chronischen muskuloskeletalen Schmerzen zu unterscheiden. Eine systematische Literaturrecherche wurde in folgenden Literatur-Datenbanken durchgeführt: Web of Knowledge (einschliesslich PubMed und Cinahl), Scopus und Cochrane. Zwei Autoren prüften die gefundenen Studien unabhängig von einander anhand der Titel und Zusammenfassungen. Studien wurden berücksichtigt, wenn diese Originaldaten von Tests enthielten, welche nicht-maximale Leistungsfähigkeitt objektiv feststellen konnten, wenn maximale Leistung von den Patienten gefordert war. Die Daten der berücksichtigten Studien wurden von zwei Autoren unabhängig von einander extrahiert. Dieselben Autoren beurteilten die Studien nach den Subskalen der COSMIN-Checkliste „criterion validity" und „hypothesis testing". Eine Ergebnissynthese wurde durchgeführt. Von sieben eingeschlossenen Studien, enthielten fünf einen Referenz-Standard, um nicht-maximale Leistungsfähigkeit festzustellen. Drei Studien waren von guter methodologischer Qualität und konnten nicht-maximale Leistungsfähigkeit mit einer Spezifität von 75% bis 100% feststellen. Zusammenfassend lässt sich sagen, dass es Evidenz dafür gibt, dass nicht-maximale Leistungsfähigkeit bei Patienten mit chronischen, unspezifischen Rückenschmerzen mittels Beobachtungskriterien im Rahmen von Tests der Evaluation der funktionellen Leistungsfähigkeit (EFL) oder mittels lumbal platzierter Bewegungsmesser identifiziert werden kann.

Kapitel 3 beschreibt eine Studie zur Überprüfung der Reliabilität (=Zuverlässigkeit) von Skalen mit funktionellen Beobachtungskriterien, die den Grad der physischen Leistung bei Tests innerhalb der Evaluation der Funktionellen Leistungsfähigkeit (EFL) beurteilen. Einundzwanzig Physiotherapeuten beurteilten unabhängig von einander, je einmal im Abstand von zehn

Monaten, 18 Videosequenzen von EFL-Tests. Die in den Videos demonstrierte physische Belastung wurde anhand einer Ordinalskala (P_{ED}) mit vier Kategorien („leicht-mässig", „schwer", „maximal", „über maximale Belastung") gemäss den funktionellen Beobachtungskriterien nach Isernhagen bewertet. Zusätzlich wurden mit einer dichotomen Skala (S_{ED}) beurteilt, ob der Patient einen submaximalen Effort erbringt und den Test vor Erreichen eines funktionellen Limits vorzeitig abbricht oder nicht. Die Übereinstimmung desselben Beurteilers im Zeitabstand von 10 Monaten (intra-rater Reliabilität) und Urteilseinigkeit zwischen der Beurteiler (inter-rater Reliabilität) wurde mit Cohen's Kappa-Koeffizienten (κ) und Prozentangaben (%) der Übereinstimmung berechnet (akzeptable Reliabilität: $\kappa > 0.60$, Übereinstimmung ≥80%). Die Kappa-Koeffizienten für die intra-rater Reliabilität aller EFL-Tests waren für die P_{ED}-Skala 0.49 und für die S_{ED}-Skala 0.68. Die Kappa-Koeffizienten für die inter-rater Reliabilität aller EFL-Tests der P_{ED}-Skala waren 0.51 und für die S_{ED}-Skala 0.72. Für die einzelnen EFL-Tests variierte die inter-rater Reliabilität der P_{ED}-Skala von schwach bis fast perfekter Übereinstimmung. Eine akzeptable Reliabilität für die einzelnen EFL-tests wurden gemessen in 38% der P_{ED}-Skala, sowie in 67% der S_{ED}-Skala. Die Reliabilität der Beobachtungskriterien ist bei EFL-Tests wie zum Beispiel Heben von Lasten höher, als bei statischen Tests wie zum Beispiel Knien oder Gehen. Die Studie kommt zum Ergebnis, dass die zweiteilige Skala zur Beurteilung des submaximalen Efforts (S_{ED}) zuverlässiger ist, als die Skala mit vier Kategorien der physischen Belastung (P_{ED}). Regelmässige Schulungen der Beurteiler könnte die Reliabilität der Beobachtungsklassifikationen für EFL-Tests weiter erhöhen.

Kapitel 4 beschreibt eine Studie, die die Reliabilität bei Studienbeginn (=Test) und nach sieben Tagen (=Retest) sowie die Sicherheit von EFL-Tests bei Patienten mit Beschwerden nach HWS-Distorsionstrauma (HD) untersuchte. Zweiunddreissig Studienteilnehmer (11 Frauen, 21 Männer, mittleres Alter 39.6 Jahre) mit HD Grad I und II nahmen an der Studie teil. Die EFL bestand aus 12 Tests: fünf Hebetests, Handkraft, wiederholtes Greifen, Arbeit über Schulterhöhe, Gehgeschwindigkeit und Dreiminutenstufentest. Die EFL-Tests dauerten ca. 60 Minuten. Die Dauer zwischen Test und Retest war im Mittel sieben Tage. Die Interklassen Korrelationen (model 1) (ICCs) und Limits of Agreement (LoA) wurden berechnet. Die Sicherheit der Teilnehmer, welche die EFL-Tests durchführten, wurde mit dem Fragebogen "Pain Response Questionnaire", mit funktionellen Beobachtungskriterien und Überprüfung der Herzfrequenz kontrolliert. ICCs variierten zwischen mässig im „Dreiminutenstufentest" bis exzellent bei „Heben horizontal". Die LoA relativ zu den gemittelten Werten beliefen sich zwischen 15% (50m Geh-Test) und 57% (Heben Taille zu Kopfhöhe). Nach einem kurzzeitigen Anstieg nach den EFL-Tests verringerten sich die Schmerzen innerhalb weniger Tage auf das Schmerzniveau vor Durchführung der EFL-Tests. Die funktionellen Beobachtungen, wie auch die Resultate aus den Herzfrequenzmessungen zeigten keine Auffälligkeiten oder Abweichungen bei den definierten Sicherheitskriterien. Die Test-Retest Reliabilität war

moderat in zwei, gut in fünf und exzellent in weiteren fünf EFL-Tests. Die Sicherheitskriterien wurden erfüllt. Die Resultate bezogen auf den einzelnen Patienten sind aufgrund der Bandbreite der LoA-Werte mit der notwenigen Umsicht zu interpretieren.

Kapitel 5 beschreibt die Konstrukt-Validität (=Gültigkeit) von EFL-Tests, bei arbeitsunfähig geschriebenen Patienten mit Beschwerden nach HWS-Distorsionstraumen (HD). Die Teilnehmer wurden aufgrund ihrem kulturellen Hintergrund bzw. der Muttersprachen (Deutsch, Nicht-Deutsch) in zwei Gruppen klassifiziert. In einer Querschnittstudie wurden 314 Teilnehmer (42% Frauen, mittleres Alter 36.7 Jahre) mit HD Grad I und II zu einem interdisziplinären Assessment zugewiesen, bei dem auch EFL-Tests durchgeführt wurden. Geprüft wurden vier EFL-Tests (Handkraft, Heben Taille- zu Kopfhöhe, Arbeit über Schulterhöhe und wiederholtes Greifen). Für die Konstrukt-Validität wurden weitere Variablen, wie die Schmerzintensität, die wahrgenommene physische Belastbarkeit, die Beeinträchtigungen im Alltag und die Angst und die Depressivität per Fragebogen erhoben. Die Konstrukt-Validität wurde wie folgt definiert: a priori wurden Hypothesen formuliert, welche mit Korrelationen den Zusammenhang zwischen EFL-Tests und den oben genannten Konstrukt-Variabeln überprüften. Die Resultate der Studie zeigen, dass die Männer im Durchschnitt eine signifikant höhere Handkraft (+17.5 kg) aufweisen und mehr Gewicht über Schulterhöhe heben (+3.7 kg). Drei von vier geschlechterbezogenen Hypothesen wurden bestätigt. Die Unterschiede bezüglich den EFL-Tests waren signifikant zwischen Patienten mit unterschiedlicher Muttersprache. Die Korrelation zwischen EFL-Tests und Schmerzintensität war tief, die Korrelation zwischen EFL-Tests und der wahrgenommenen, physischen Belastbarkeit war moderat. Die Korrelation zwischen EFL-Tests und Alltagseinschränkungen, sowie zwischen EFL-Test und selbsteingeschätzter Angst- und Depressivität war tief. Die beiden Gruppen mit unterschiedlicher Muttersprache zeigten sehr ähnliche Korrelationskoeffizienten zwischen EFL-Tests und Konstrukt-Variablen. Die Mehrheit der a-priori gestellten Hypothesen wurde nicht verworfen. Die Schlussfolgerung der Studie ist: die Konstrukt-Validität der EFL-Tests ist gut bei arbeitsunfähig geschriebenen Patienten nach HD und unterschiedlicher Muttersprache. Das Resultat sollten in anderen Studien verifiziert werden.

Die Studie in **Kapitel 6** untersucht die Eignung von EFL-Tests, um die zukünftige Arbeitsfähigkeit vorauszusagen. Eine prospektive Kohortenstudie wurde im Rahmen des ambulanten Assessments der Arbeitsorientierten Rehabilitation in der Rehaklinik Bellikon durchgeführt. Die Studien-Teilnehmer waren arbeitsunfähige Patienten mit HD, 6 bis 12 Wochen nach Unfall. Diese Patienten haben acht unterschiedliche EFL-Tests durchgeführt. Die Ziel-Variable (=Outcome) war die Arbeitsfähigkeit (AF) in %, welche zu Studienbeginn, sowie nach 1, 3, 6, und 12 Monaten erhoben wurde. Korrelationen zwischen FCE-Tests und AF wurden berechnet. Eine lineare Multilevel-Analyse wurde durchgeführt, um den Zusammenhang

zwischen den EFL-Tests und der künftigen AF zu berechnen. Insgesamt nahmen 267 Patienten mit Beschwerden nach HD Grad I und II an der Studie teil. Die mittlere AF stieg von 20.8% zu Beginn, auf 83.2% nach 12 Monaten. Die Korrelationen zwischen EFL-Tests und AF war gering nach 12 Monaten bezüglich Test „Heben Boden zu Taillenhöhe" und schwach für den Test „Gehen" nach 3 Monaten. Die Korrelation zwischen EFL-Tests und AF nahm im Verlauf der 12 Monate ab. Die künftige AF wurde von EFL-Tests nicht vorausgesagt. Variablen, welche die AF voraussagen, waren: die Dauer seit Unfall (β=23.74), die Muttersprache des Patienten (β=5.49), die AF bei Beginn der Testung (β=1.01) und die selbst eingeschätzten Alltagseinschränkungen (β=-0.20). Die Ergebnisse der Studie weisen darauf hin, dass EFL-Tests bei Beginn mit der AF assoziiert sind, jedoch die künftige AF nicht voraussagen. Der Dauer seit Unfall, die Muttersprache, die AF bei Beginn und die Alltagseinschränkungen können die künftige AF voraussagen.

Kapitel 7 beschreibt eine Studie zur Überprüfung der Gütekriterien des PACT-Fragebogens[2] (Performance Assessment and Capacity Testing). Der PACT-Fragebogen umfasst 50 Fragen. Jede Frage ist mit einem passenden Bild versehen. Der PACT-Fragebogen misst, die selbsteingeschätzte körperliche Leistungsfähigkeit bei Tätigkeiten, die die Wirbelsäule belasten. Dreihundertvier Patienten mit HD Grad I und II aus dem ambulanten Assessments der Arbeitsorientierten Rehabilitation der Rehaklinik Bellikon nahmen teil. Die interne Konsistenz (=Stabilität) wurde mit dem Cronbach's α Koeffizienten untersucht. Die Konstrukt-Validität wurde mit 8 a priori Hypothesen überprüft. Die strukturelle Validität wurde mit der Hauptkomponenten-Analyse (principal component analysis, PCA) untersucht. Die Reliabilität wurde mit zwei aufeinander folgenden Messungen im Abstand von ca. 7 Tagen durchgeführt (Test und Retest). Für die Test-Retest Reliabilität wurden Daten von 32 Patienten mittels Interklassen Korrelationen (ICCs) und Limits of Agreement (LoA) analysiert. Die prädiktive Validität für die künftige Arbeitsfähigkeit (AF) in % nach 1, 3, 6 und 12 Monaten wurde bestimmt durch die Area Under the Curve (AUC, >0.70). Die Outcome variable Nicht-Rückkehr zu Arbeit (N-RA) wurde anhand von zwei Schwellenwerten berechnet: AF<50% und AF<100%. Die Resultate der Studie zeigen, dass der Anteil der Patienten mit N-RA sich innerhalb von 12 Monaten von 50% auf 14% verringerte. Cronbach's α war >0.90. PCA-Resultate wiesen auf eine Eindimensionalität hin. ICC-Werte betrugen 0.86 (95% CI: 0.71; 0.93), die LoA waren ±33 Punkte (von max. 200 Punkte). Sieben von 8 Hypothesen zur Konstrukt-Validität wurden nicht verworfen. Die AUC verringerte sich im Verlauf der Messungen von 0.71 im 1. Monat auf 0.68 im 12 Monat für den Schwellenwert AF<50%. Für den Schwellenwert AF<100% waren die Werte 0.71 im 1. Monat und 0.59 im 12. Monat. Zusammenfassend lässt sich sagen, dass die Stabilität, die Zuverlässigkeit, die Interne Konsistenz und die Konstrukt- und Struktur-Validität

2 Im englischen Sprachraum ist der PACT-Fragebogen unter dem Namen Spinal Function Sort, kurz SFS, bekannt.

des PACTs bei Patienten mit sub-akuten Nackenbeschwerden nach HD ausreichend sind. Die LoA-Werte sind gross. Kurzfristig konnte die N-RA bis 1 Monat mittels PACT-Fragebogen vorausgesagt werden, jedoch nicht über einen längeren Zeitraum von 12 Monaten.

Die Studie in **Kapitel 8** beschreibt die Übersetzung, die transkulturellen Adaptation und die Validierung des PACT-Fragebogens in französischer (SFS-F) und deutscher (SFS-G) Sprache. Es wird angenommen, dass der PACT die selbsteingeschätzte, körperliche Leistungsfähigkeit misst. Dreihundertvierundvierzig Patienten mit Rückenbeschwerden aus zwei Rehabilitations-Kliniken, aus der französischsprachigen (n=87) und der deutschsprachigen Schweiz (n=257), nahmen an der Studie teil. Die Konstrukt-Validität wurde mittels Korrelationen aus der physischen und mentalen Dimension des SF-36 Fragebogens berechnet. Zusätzlich wurden Korrelationen der Konstrukt-Validität mittels visueller Schmerzskala (VAS) und der Selbstbeurteilung von Angst und Depressivität (HADS) analysiert. Die interne Konsistenz (Stability) wurde mittels Cronbach's α berechnet. Die Test-Retest Reliability wurde mit 65 zusätzlichen Probanden durchgeführt und mittels Korrelation Koeffizienten (ICC) und Limits of Agreement (LoA) berechnet. Für die beiden Versionen des PACTs wurden Cronbach's α und ICC von >0.90 gemessen. Die Korrelationen waren mit der SF-36 Dimension Physical Functioning moderater für beide SFS-Sprachversionen. Die Korrelationen mit selbstbeurteiltem Schmerz waren gering für den französischsprachigen und moderat für den deutschsprachigen PACT. Die Korrelationen mit Angst und Depressivität war für beide Sprachversionen des PACTs gering bis nicht vorhanden. Die Reliabilität war gut (ICC>0.90) für beide Sprachversionen des PACTs. LoA waren -11.5 und 12.1 Punkte für den französischsprachigen PACT , und -27.7 sowie 30.2 Punkte für den deutschsprachigen PACT. Die Konvergente-Validität war gut in den „Physical Scales des SF-36 und war moderat mit visueller Schmerzskala (VAS). Die divergente Validität war tief im Vergleich mit dem „mental scale" des SF-36 in beiden Sprachversionen und tief mit HADS für den französischsprachigen PACT (moderat für den deutschsprachigen PACT). Einen wesentlichen Unterschied gab es zwischen den LoA-Werten der beiden Sprachversionen des PACTs, wobei die LoA-Werte des deutschsprachigen PACTs mehr als doppelt so gross waren. Beide Versionen, die deutschsprachige und französischsprachige, erscheinen valide und zuverlässig bezüglich der Bewertung der selbsteingeschätzten körperliche Leistungsfähigkeit bei Patienten mit Rückenbeschwerden.

Kapitel 9 diskutiert die Hauptergebnisse sowie die Stärken und Schwächen dieser Dissertation anhand von methodologischen Gesichtspunkten. Die Bedeutung der Resultate dieser Dissertation wird für Patienten, EFL-Anwender und Zuweiser erläutert. Aufgeführt werden Empfehlungen für weitere Forschung. Im Abschnitt „Valorization" (Nutzen) werden die konkrete Umsetzung der Ergebnisse im klinischen Alltag beschrieben und mögliche nächste Schritte diskutiert. Die Schlussfolgerungen dieser Dissertation werden am Ende des Kapitels erläutert.

Diese Dissertation hat die Gütekriterien von EFL-Tests und des PACT-Fragebogens umfassend untersucht. Es zeigte sich, dass EFL-Tests valide (=gültig) sind, um das Leistungsverhalten (Effort) von Patienten mit Rückenschmerzen zu erfassen. Die Reliabilität der EFL-Anwender, die mit standardisierten funktionellen Beobachtungskriterien die Leistungsfähigkeit bewerten, war gut bei den Hebe- und Tragetests, aber ungenügend für statische Tests und Gehtests. Bezüglich Reliabilität war die Skala mit zwei Kategorien, deutlich besser als die Reliabilität der Skala mit vier Kategorien. EFL-Tests zeigten akzeptable Werte hinsichtlich der Test-Retest-Reliabilität und der Validität bei Patienten mit HD Grad I und II mit unterschiedlichen Muttersprachen. Bezogen auf den individuellen Patienten sollte bei der Anwendung der EFL-Tests und des PACT-Fragebogens die Bandbreite des Messfehlers berücksichtigt werden. Die zukünftige Arbeitsfähigkeit kann mit EFL-Tests bei Patienten mit sub-akuten HD nicht vorausgesagt werden. Hingegen scheinen die Dauer seit Unfall, die aktuelle Arbeitsfähigkeit, die Muttersprache und die selbst eingeschätzte Alltagseinschränkung die zukünftige Arbeitsfähigkeit vorauszusagen. Da es sich dabei um nicht modifizierbare Faktoren handelt, ist der zielgerichtete Einsatz von rehabilitativen Massnahmen begrenzt, die eine raschere Rückkehr zur Arbeit ermöglichten. Die französische und deutsche Version des PACT-Fragebogens sind reliabel und valide bei Patienten mit Rückenbeschwerden oder HD Grad I und II. Der PACT-Fragebogen kann die zukünftige Arbeitsfähigkeit von Patienten mit sub-akuten HD Grad I und II nicht voraussagen. Diese Resultate weichen somit von Ergebnissen einer früheren Studie mit Patienten mit chronischen Kreuzbeschwerden ab. Es werden Vorschläge gemacht, wie der PACT-Fragebogen verbessert werden z.B. indem man die Anzahl der Fragen um die Hälfte reduziert, ohne dass sich die Messeigenschaften verschlechtern.

Der Autor wünscht, dass die Ergebnisse dieser Dissertation Kliniker und Forscher inspiriert, diese Studien in anderen Institutionen zu wiederholen und die untersuchten Messinstrumente weiterzuentwickeln. Es ist zu hoffen, dass diese Dissertation das Bewusstsein schärft für den Einfluss der individuellen patientenbezogenen Faktoren und der Umgebungsfaktoren auf die Resultate der EFL-Tests. Mögen die Erkenntnisse dieser Dissertation Patienten, Anwender und Zuweiser beim Einsatz von EFL-Tests und PACT-Fragebogen unterstützen, damit die Wiedereingliederung von Patienten in den Arbeitsalltag weiter verbessert werden kann.

Dankwoord
Acknowledgements

"Het is helaas in Nederland nog een zeldzaamheid dat een fysiotherapeut promoveert." [Freely interpreted: it's an oddity in the Netherlands that a physiotherapist does a promotion], cited from the introduction of the thesis of Pieter U. Dijkstra, 1993. This statement applies (still) to the situation in Switzerland. Therefore, I feel privileged and owe gratitude to many persons without whom this thesis would never have been possible.

First and foremost, I must thank my wonderful family, including my wife Corinne, my children Emma, Sofia, and Teo, my parents, Giammario and Myrta, and my brothers, Ivar and Francesco. Any accomplishments on my account, including this thesis, are due to your ceaseless support. Corinne, you have always had faith in my endeavors and I owe it all to you.

I would like to thank Michael Oliveri, past-head of the Department of Work Rehabilitation at the Rehaklinik Bellikon. If Functional Capacity Evaluation (FCE) and work-rehabilitation are known in the Swiss Rehabilitation and Insurance field, it's because of Michael. He pushed the field with his great passion, innovation and endurance. Dear Michael, you encouraged me, gave me confidence and the freedom to perform this thesis within my full-time position at the department.

I met my first supervisor, Prof. Dr. Michiel Reneman, in 2008 in Liverpool at the IASP satellite symposium, while singing *"O sole mio"* with Martin and Roberto. Michiel, you did not hesitate a second to support the idea of this thesis. *Hartstikke bedankt*, for your motivational attitude during the thesis-trajectory, your vast knowledge on FCE, and your excellent networking skills. We will never forget the tasty *hete bliksem* (traditional potato dish) you cooked at our home in Switzerland!

Prof. Dr. Pieter Dijkstra, co-supervisor, has been a great mentor to me, as he has been for many other young researchers during his impressive research career. Dear Pieter, after you returned the manuscripts, it has become clear to me what the term "thorough review" means. The stats-sessions at your office with fresh pâté and bread followed by Swiss chocolate will remain unforgettable. *Hartelijke dank*, for your assistance during the many cooking sessions and the guided *fiets*-tours in the Groningen county.

Prof. Dr. Jan Geertzen, co-supervisor, has EXPANDed the Department of Rehabilitation Medicine in Groningen to be probably one of the world's largest in this research field. *Beste Jan, h-stikke bedankt* for giving me the opportunity to embed this PhD thesis at your Department. Your manuscript reviews were always fast, concise and you did not miss a "dot".

Special thanks to my colleague from the management-team at the Department of Work Rehabilitation, Rehaklinik Bellikon, Sandra Hedinger, and to the team-leaders Grit Hoffmann, Yves Weder, Nicole Saghy and Jonas Bühler who had to take on extra work when I was in the Netherlands or busy with this thesis.

Additional thanks to the many physiotherapists and physicians of the Department of Work Rehabilitation who provided invaluable help in the assessment of patients. Vielen Dank für eure Unterstützung!

Besonders bedanken möchte ich mich bei folgenden Kollegen aus Bellikon: Peter Erhart, Claudia Diethelm, Axel Gehrke und Linda Fröhli. Sie waren mir bei der Eingabe und Zusammenführung der diversen Daten eine sehr grosse Unterstützung.

Grossen Dank gebührt den verantwortlichen Personen der suva (Schweizerische Versicherungsanstalt), für die finanzielle Unterstützung des Forschungsprojektes und das zugesprochene Vertrauen. Speziell erwähnt seien Thomas Mäder, Felix Weber, Rita Schaumann-von Stosch, Christian Ludwig, René Meier und die Mitglieder der Forschungskommission MSK-005.

Herzlich bedanken möchte ich mich bei Prof. Dr. Achim Elfering. Lieber Achim, seit unserer ersten Begegnung am Clinical Research Forum im Jahr 2005 hat mich der Gedankenaustausch mit dir immer wieder inspiriert. Danke, dass du meine Teilnahme an Kursen der Graduate Health School unterstützt hast.

De leden van de beoordelingscommissie, Prof. Dr. Ute Bültmann, Prof. Dr. Bart Koes en Prof. Dr. Rob Smeets, dank ik voor het zorgvulig lezen en beordelen van dit proefschrift.

I am grateful to be supported by the *paranimfen* Trix Jansen and Haitze de Vries. Trix is an ergonomist and has studied physiotherapy in the Netherlands. She was one of the first professionals at the Department of Work Rehabilitation. For me, Trix, you are probably one the most multitalented health care providers in the field, covering various roles as expert clinician, manager and researcher. I have learned a lot from you, *hartelijk bedankt*! I met Haitze, the real *Frisian*, when he was a PhD student at the UMCG. With Haitze, I share many common interests such as rehabilitation of workers – he was for many years at the Spine center in Rotterdam – skating, cycling or gardening. Beste Haitze, thank you for helping me in organizing the thesis defense.

Tante grazie to the wives of my supervisors, Judith, Bernarda and José. They have allowed me to use "their" kitchens during the many cooking sessions and dinners. My apologies, if the kitchen was not as clean as it used to be before my cooking sessions.

Among others, two extraordinary personalities in the field of FCE and work rehabilitation have inspired my research. Susan Isernhagen (Duluth, USA) has thrived for 30 years in the field of FCE. She has encouraged many young researchers to improve the evidence base of FCE. Dear Sue, during your visit to Switzerland, you impressed me with your updated knowledge and your enthusiasm. Assoc. Prof. Doug Gross (Edmonton, Canada) has delineated the FCE with methodologically "robust" studies. Dear Doug, I hope that the curvy bus drive up to Valens has shaken your brain cells stimulating further innovative studies. Go for it!

During my stays in Groningen and Haren, I was warmly welcomed and hosted by many kind persons. *Dank u wel*! to Bert Eissens, Wim Jorritsma, Rita Schiphorst, Berry van Holland, Ben Evers, Remko Soer, Franka Waterschoot, Christa Nanninga, Clemens Rommers, Katy van Hasselt (location Haren), Assoc. Prof. Dr. Sandra Brouwer, Prof. Dr. Corry van der Sluis, Sietke Postema, Raoul Bongers, Roy Stewart, Priya Vart, Jan Buitenhuis, Jerry Spanjer, Prof. Dr. Klaas Postema, Inge Zweerts de Jong-Veeze (location Groningen) and Sandra Lakke (Hanzehogeschool Groningen) among many others.

Many contacts, visits and invited lectures in the Netherlands were packed with inspiring encounters. I would like to explicitly express my profound gratitude to all of them: Assoc. Prof. Dr. Paul Kuijer, Jan Hoving, Prof. Dr. Haije Wind, Vincent Gouttebarge (all Academic Medical Center, Amsterdam), Martijn Heymans, Iris Eckhout, Prof. Dr. Jos Twisk, Tobias van den Berg, Luiz Hespanhol Jr. (all Vrije Universteit, Amsterdam), Martin van der Esch (Reade Rehabilitation center, Amsterdam), Prof. Dr. Rob Smeets, Ivan Huijnen, Prof. Dr. Janine Verbunt, Hanne Kindermans, Carolien Bastiaenen, Judith Sieben (all Maastricht) and Albére Köke (Hoensbroek).

Finally, the power of individuals who contributed to the advances in the field rehabilitation sciences in Switzerland deserves great respect. Many of these individuals became good friends and colleagues: Roberto Brioschi, Felix Angst, Thomas Benz (Bad Zurzach), Lorenz Radlinger, Gere Luder, Martin Verra (all Bern), Dominique Monnin, Lara Allet (both Geneva), Roger Hilfiker (Leukerbad), Luca Scascighini (Lugano), Corina Schuster (Rheinfelden), Jan Kool (Valens), Markus Wirz, André Meichtry, Omega Erika Huber, Prof. Dr. Astrid Schämann, Prof. Dr. Karin Niedermann (all Winterthur), Ruud Knols and Jaap Swanenburg (both Zurich).

In case I might have forgotten to acknowledge somebody here, my apologies and thank you very much.

About the author

Maurizio Alen Trippolini was born 1975 in Samedan, in Rhaeto-Romanic-speaking mountain area of Engadine-St. Moritz. After completing gymnasium at the Lyceum Alpinum Zuoz and serving the Swiss Army he worked as a professional snow sports instructor. Maurizio acted as Head of Education in snowboarding for several years. In 1996 he started his studies in physiotherapy at the University of Applied Sciences in Southern Switzerland (SUPSI) in Landquart, he graduated in 2000. In 2005 he finished his Master of Physiotherapy Sciences at the Department of Epidemiology, Faculty of Health Sciences at the University of Maastricht, The Netherlands. In March 2011, he became a PhD student at the Department of Rehabilitation Medicine at the University of Groningen, The Netherlands.

Maurizio started his professional career as a full-time clinical physiotherapist working in orthopedic and musculoskeletal departments in general hospitals. From 2002 till 2006 he worked as a physiotherapist in the Department of Ergonomics at the Rehaklinik Bellikon, an enterprise of the Swiss Accident Insurance Fund (SUVA). At the department, Maurizio worked under the supervision of Michael Olivieri, the Swiss pioneer of work rehabilitation, as part of an interdisciplinary team of psychologists, physicians, vocational trainers and case managers, coaching injured workers with chronic musculoskeletal pain back to work. Other elements of his role were the evaluation of functioning using Functional Capacity Evaluation and guiding patients through progressive return-to-work programs. In 2006 he was appointed to a management position in the Department of Work Rehabilitation at the Rehaklinik Bellikon. He was involved in expanding the department which has grown from one to four interdisciplinary teams.

Maurizio supervises Bachelor and Master physiotherapy students. He works as a lecturer and course coordinator at the postgraduate study centre in Bellikon for courses on rehabilitation of patients with whiplash injuries, strength training and other courses. Additionally, he is the coordinator for the post-graduate course in work rehabilitation at SUPSI. Maurizio co-founded the annual Clinical Research Forum in 2005 and continues to be an organizer for it. He is a member of the board of various research funds including the Swiss Physiotherapy Sciences Foundation. Currently Maurizio combines management of the department with patient care, teaching and clinical research. He spends his leisure time with his family, enjoying cycling, snow sports, and mountaineering and he is a volunteer fire-fighter.

Maurizio lives with his wife and three children near Zurich, Switzerland.

Research Institute SHARE / previous dissertations

RESEARCH INSTITUTE SHARE

This thesis is published within the **Research Institute SHARE** of the University Medical Center Groningen / University of Groningen.

Further information regarding the institute and its research can be obtained from our internetsite: www.rug.nl/share.

More recent theses can be found in the list below.
((co-) supervisors are between brackets)

2014

Suwantika AA
Economic evaluations of non-traditional vaccinations in middle-income countries: Indonesia as a reference case
(prof MJ Postma, dr K Lestari)

Behanova M
Area- and individual-level socioeconomic differences in health and health-risk behaviours; a comparison of Slovak and Dutch cities
(prof SA Reijneveld, dr JP van Dijk, dr I Rajnicova-Nagyova, dr Z Katreniakova)

Dekker H
Teaching and learning professionalism in medical education
(prof J Cohen-Schotanus, prof T van der Molen, prof JW Snoek)

Dontje ML
Daily physical activity in patients with a chronic disease
(prof CP van der Schans, prof RP Stolk)

Gefenaite G
Newly introduced vaccines; effectiveness and determinants of acceptance
(prof E Hak, prof RP Stolk)

Dagan M
The role of spousal supportive behaviors in couples' adaptation to colorectal cancer
(prof M Hagedoorn, prof R Sanderman)

Monteiro SP
Driving-impairing medicines and traffic safety; patients'perspectives
(prof JJ de Gier, dr L van Dijk)

Bredeweg S
Running related injuries
(prof JHB Geertzen, dr J Zwerver)

Mahmood SI
Selection of medical students and their specialty choices
(prof JCC Borleffs, dr RA Tio)

Krieke JAJ van der
Patients' in the driver's seat; a role for e-mental health?
(prof P de Jonge, prof M Aielllo, dr S Sytema, dr A Wunderink)

Jong LD de
Contractures and hypertonia of the arm after stroke; development, assessment and treatment
(prof K Postema, prof PU Dijkstra)

Tiessen AH
Cardiovascular risk management in general practice
(prof K van der Meer, prof AJ Smit, dr J Broer)

Bodde MI
Complex Regional Pain Syndrome type 1 & amputation
(prof JHB Geertzen, prof PU Dijkstra, dr WFA van der Dunnen)

Lakke AE
Work capacity of patients with chronic musculoskeletal pain
(prof JHB Geertzen, prof MF Reneman, prof CP van der Schans)

Silarova B
Unraveling the role of sense of coherence in coronary heart disease patients
(prof SA Reijneveld, dr JP van Dijk, dr I Rajnicova-Nagyova)

Weening-Dijksterhuis E
Physical exercise to improve or maintain Activities of Daily Living performance in frail institutionalized older persons
(prof CP van der Schans, prof JPJ Slaets, dr MHG de Greef, dr W Krijnen)

Koolhaas W
Sustainable employability of ageing workers; the development of an intervention
(prof JJL van der Klink, prof JW Groothoff, dr S Brouwer)

Flach PA
Sick leave management beyond return to work
(prof JW Groothoff, prof U Bültmann)

2013

Bosker BH
Pitfalls in traditional and innovative hip replacement surgery
(prof SK Bulstra, dr CCPM Verheyen, dr HB Ettema)

Holwerda A
Work outcome in young adults with disabilities
(prof JJL van der Klink, prof JW Groothoff)

Mohseninejad L
Uncertainty in economic evaluations: implications for healthcare decisions
(prof E Buskens, dr TL Feenstra)

Cornelius LR
A view beyond the horizon; a prospective cohort study on mental health and long-term disability
(prof JJL van der Klink, prof JW Groothoff, dr S Brouwer)

Sobhani S
Rocker shoes for ankle and foot overuse injuries: a biomechanical and physiological evaluation
(prof K Postema, prof ER van den Heuvel)

Pitel L
Sociocultural determinants, gender and health-related behaviour in adolescence
(prof SA Reijneveld, dr JP van Dijk, dr A Madarasova-Geckova)

Majerníková M
Sef-rated health and mortality after kidney transplantation
(prof JW Groothoff, dr JP van Dijk, dr J Rosenberger, dr R Roland)

Verschuren J
Sexuality and limb amputation: perspectives of patients, partners and professionals
(prof JHB Geertzen, prof PU Dijkstra, prof P Enzlin)

Riphagen-Dalhuisen J
Influenza vaccination of health care workers
(prof E Hak)

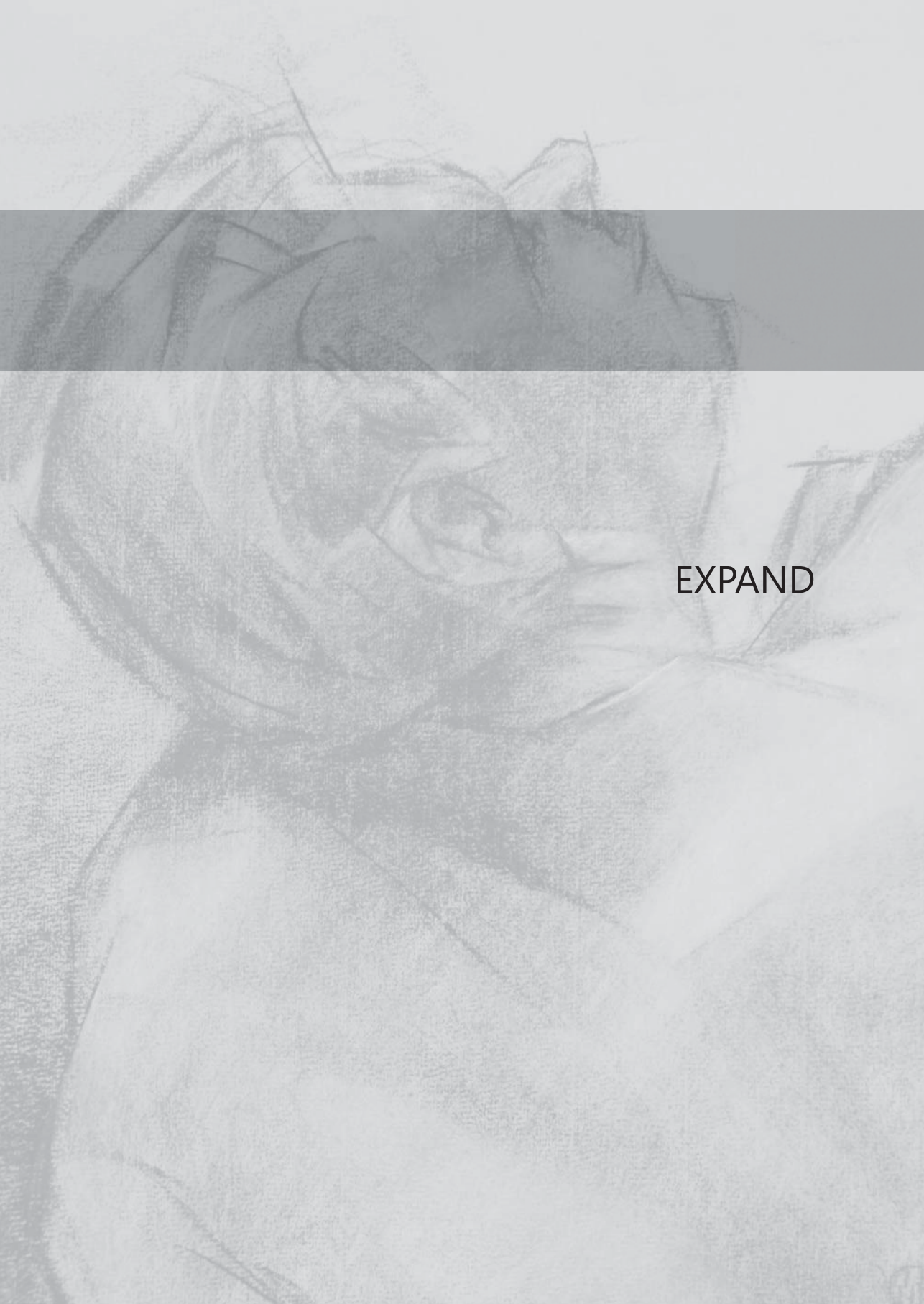
Hasselt FM van
Improving the physical health of people with severe mental illness; the need for tailor made care and uniform evaluation of interventions
(prof AJM Loonen, prof MJ Postma, dr MJT Oud, dr PFM Krabbe)

Piening S
Communicating risk effectively
(prof FM Haaijer-Ruskamp, prof PA de Graeff, dr PGM Mol, dr SMJM Straus)

Siebelink MJ
The child as a donor; a multidisciplinary approach
(prof HBM van de Wiel, prof PF Roodbol)

Sidorenkov G
Predictive value of treatment quality indicators on outcomes in patients with diabetes
(prof FM Haaijer-Ruskamp, prof D de Zeeuw)

For more 2013 and earlier theses visit our website.


EXPAND

**Wetenschappelijk onderzoek afdeling Revalidatiegeneeskunde –
Centrum voor Revalidatie UMCG**

EXPAND

Extremities, Pain and Disability

Missie: EXPAND draagt bij aan participatie en kwaliteit van leven van mensen met aandoeningen en amputaties van de extremiteiten of met pijn aan het bewegingsapparaat.

EXPAND omvat twee speerpunten: onderzoek naar aandoeningen aan en amputaties van extremiteiten met nadruk op stoornissen, activiteiten en participatie en onderzoek naar chronische pijn en arbeidsparticipatie. EXPAND draagt bij aan het UMCG-brede thema Healthy Ageing.

**Research Department of Rehabilitation Medicine –
Center for Rehabilitation UMCG**

EXPAND

Extremities, Pain and Disability

Mission: EXPAND contributes to participation and quality of life of people with conditions and amputations of the extremities and musculoskeletal pain.

EXPAND focuses on two spearheads: research on the conditions and amputations of the extremities with emphasis on body functions and structures, activities and participations, and chronic pain and work participation. EXPAND contributes to Healthy Aging, the focus of the UMCG.